

Proceedings

Revised September 7, 2012

Copyright © HOT CHIPS, 2012. All rights reserved. All trademarks property of their respective owners.

HOT CHIPS 24

Conference Day 1
Tuesday, August 28, 2012

8:30 – 8:45am	**Opening Remarks**	Christos Kozyrakis & Rumi Zahir	
8:45 – 10:15am	**Session 1**	**Microprocessors**	
	Power Management of the Third Generation Intel Core Micro Architecture formerly codenamed Ivy Bridge	Sanjeev Jahagirdar, **Intel**	
	AMD's "Jaguar": A next generation low power x86 core	Jeff Rupley, **AMD**	
	proAptiv: Efficient Performance on a Fully-Synthesizable Core	Ranganathan Sudhaka, **MIPS**	
10:15 – 10:45am	**Break**	**Student poster presentations**	
10:45 – 12:15pm	**Session 2**	**Fabrics & Interconnects**	
	Swizzle Switch: A Self-Arbitrating High-Radix Crossbar for NoC Systems	Ronald Dreslinski**, Michigan**	
	FPGA Augmented ASICs: The Time Has Come	David Riddoch, **Solarflare**	
	SwitchX Virtual Protocol Interconnect (VPI) Switch Architecture	Diego Crupnicoff, **Mellanox**	
12:15 – 1:30pm	**Lunch**		
1:30 – 2:30pm	**Keynote 1**	**The Surround Computer Era**	
	Mark Papermaster, CTO, **AMD**		
2:30 – 2:50pm	**Break**	**Student poster presentations**	
2:50 – 4:20pm	**Session 3**	**Many Core and GPU**	
	AMD Radeon HD7970 Graphics Core Next (GCN) Architecture	Michael Mantor, **AMD**	
	AMD "Trinity" APU		
	Intel® Many Integrated Core Architecture -The first Intel® Xeon Phi™ coprocessor (codename Knights Corner)	Sebstian Nussbaum, **AMD** George Chrysos, **Intel**	
4:20 – 4:50pm	**Break**	**Student poster presentations**	
4:50 – 5:50pm	**Session 4**	**Multimedia & Imaging**	
	ADI's Revolutionary BF60x Vision Focused Digital Signal Processor System On Chip : 25 Billion Operations/Sec @ 80 mW and Zero Bandwidth	Robert Bushey, **ADI**	
	Visconti2 – A Heterogeneous Multi-Core SoC for Image-Recognition Applications	Masato Uchiyama, **Toshiba**	
5:50 – 6:50pm	**Session 5**	**Integration**	
	Centip3De: A 64-Core, 3D Stacked, Near-Threshold System	Ronald Dreslinski, **Michigan**	
	FPGAs with 28Gbps Transceivers Built with Heterogeneous Stacked-Silicon Interconnects	Ephrem Wu, Suresh Ramalingam, **Xilinx**	
6:50 – 8:05pm	**Dinner**		
8:55 – 9:05pm	**Keynote 2**	**The Future of Wireless Networking**	
	Marcus Weldon, CTO, **Alcatel-Lucent**		

IEEE computer society

IEEE TCMM Technical Committee on Microprocessors and Microcomputers

acm In-Cooperation

SSCS

Copyright © HOT CHIPS, 2012. All rights reserved. All trademarks property of their respective owners.

HOT CHIPS 24

Conference Day 2
Wednesday, August 29, 2012

8:45 – 10:15am	**Session 6**	**Technology & Scalability**
	Floating-Point Processing using FPGAs An IA-32 Processor with Wide Voltage Operating Range in 32nm CMOS Reducing Transistor Variability For High Performance Low Power Chips	Michael Parker, **Altera** Gregory Ruhl, **Intel** Robert Rogenmoser, **SuVolta**
10:15- 10:45am	**Break**	**Student poster presentations**
10:45 – 12:15pm	**Session 7**	**SOC**
	High performance and efficient single-chip small cell base station SoC FSM™ (Femtocell Station Modem) – A highly integrated, performance driven, chipset solution for the small cell market Medfield Smartphone SOC – Intel's ATOM Z2460 Processor	Kin-Yip Liu**, Cavium** Luca Blessent, **Qualcomm** Rumi Zahir, **Intel**
12:15 – 1:30pm	**Lunch**	
1:30 – 2:30pm	**Keynote 3**	**Cloud Transforms IT, Big Data Transforms Business**
	Pat Gelsinger, COO Infrastructure Products,**EMC** [now CEO, **VMWare**]	
2:30 – 2:50pm	**Break**	**Student poster presentations**
2:50 – 4:20pm	**Session 8**	**Data Center Chips**
	POWER7+™: IBM's Next Generation POWER Microprocessor The Intel® Xeon® Processor E5 Family Architecture, Power Efficiency, and Performance X-Gene™: 64-bit ARM CPU and SoC	Scott Taylor, **IBM** Jeff Gilbert, Mark Rowland, **Intel** Gaurav Singh, Greg Favor, **AMCC**
4:20 – 4:50pm	**Break**	**Student poster presentations**
4:50 – 6:20pm	**Session 9**	**Big Iron**
	SPARC64 X; Fujitsu's new generation 16 core processor for the next generation UNIX servers SPARC T5: 16-core CMT Processor with Glueless 1-Hop Scaling to 8-Sockets IBM zNext: the 3rd Generation High Frequency Microprocessor Chip	Takumi Maruyama, **Fujitsu** Sebastian Turullols, Ram Sivaramakrishnan, **Oracle** Chung-Lung (Kevin) Shum, , **IBM**
6:20 – 6:30pm	**Closing Remarks**	

Copyright © HOT CHIPS, 2012. All rights reserved. All trademarks property of their respective owners.

HOT CHIPS 24	POSTERS On exhibit during Breaks
High Performance State Retention with Power Gating applied to CPU subsystems – design approaches and silicon evaluation	David Flynn, Fellow, R&D **ARM Ltd**, Cambridge, UK
Prototyping the DySER Specialization Architecture with OpenSPARC	Jesse Benson, Ryan Cofell, Chris Frericks, Venkatraman Govindaraju, Chen-Han Ho, Zachary Marzec, Tony Nowatzki, Karu Sankaralingam **University of Wisconsin-Madison**
Low Power and High Performance 3-D Multimedia Platform	Po-Han Huang, Chi-Hung Lin, Hsien-Ching Hsieh, Huang-Lun Lin and Shing-Wu Tung **Information and Communications Research Lab. Industrial Technology Research Institute**
The Model Is Not Enough: Understanding Energy Consumption in Mobile Devices	James Bornholt, **Australian National University**, Todd Mytkowicz**, Microsoft Research**, Kathryn S. McKinley**, Microsoft Research**
Efficient, Precise-Restartable Program Execution on Future Multicores	Gagan Gupta, Srinath Sridharan, and Gurindar S. Sohi, **University of Wisconsin-Madison**

Copyright © HOT CHIPS, 2012. All rights reserved. All trademarks property of their respective owners.

HOT CHIPS
August 27 - 29, 2012
Flint Center, Cupertino, CA
A Symposium on High Performance Chips
Sponsored by the IEEE Technical Committe on Microprocessors and Microcomputers in Cooperation with ACM SIGARCH
24

Welcome to Hot Chips 24

Christos Kozyrakis & Rumi Zahir

Program Committee Co-Chairs

Program Committee



- Forest Baskett, NEA
- Pradeep Dubey, Intel
- Bob Felderman, Google
- Krisztian Flautner, ARM
- Anwar Ghuloum NVIDIA
- Christos Kozyrakis, Stanford
- Chuck Moore, AMD
- Sameer Nanavati, Qualcomm
- Don Newell
- Kunle Olukotun, Stanford
- Mitsuo Saito, Toshiba
- Alan Smith, UC Berkeley
- Guri Sohi, U. of Wisconsin
- Dean Tullsen, UCSD
- Rich Uhlig, Intel
- Fred Weber
- Rumi Zahir, Intel

Program Statistics

- 54 submissions
 - Each submission was reviewed by all PC members
- 25 accepted talks
 - High-end & low-power cores, many core, graphics, server chips, multimedia SoCs, networking, …
- 5 posters
 - Poster session during morning & afternoon breaks

Keynotes

- Marc Papermaster, CTO, AMD
 - The Surround Computing Era
 - Tuesday 8/28th, 1.30pm
- Marcus Weldon, CTO, Alcatel-Lucent
 - The Future of Wireless Networking
 - Tuesday 8/28th, 8pm
- Pat Gelsinger, COO, EMC
 - Cloud Transforms IT, Big Data Transforms Business
 - Wednesday 8/29th, 1.30pm

Tutorials

- The Evolution of Mobile SoC Programming
 - Organized by Niel Trevett
 - Khronos, ArcSoft, eyeSight, Metaio, Sensor Platforms, the 11ers
- Die Stacking
 - Organized by Liam Madden
 - AMD, Amkor, Qualcomm, UMC, Xilinx

Proceedings

- For registered attendees
 - Talks, posters, and tutorials available on USB key
 - Also available online for tablet users (http://hc24.local)
- Updated talks available online after the conference
 - Including keynote talks and videos of talks
- Conference archives available online
 - http://www.hotchips.org

Conference Etiquette

- Silence your cellphones during sessions
- Question on technical talks
 - Wait until the end of the talk
 - Come to the microphone & start with name and affiliation
 - Stick to technical questions please
 - If there is a line, ask a single question
- For speakers: during the break before your talk
 - Introduce yourself to your session chair
 - Test your slides

In Remembrance



Chuck Moore



John Nickolls

Enjoy Hot Chips 24





On behalf of the Program Committee, we are pleased to welcome you to the 24th Annual Hot Chips Symposium.

We received fifty-four (54) submissions this year that covered nearly all areas of the semiconductor and computing systems industry. The seventeen-member committee carefully reviewed all submissions and selected the top twenty-five (25) that best represent the breadth and depth of our field. We also selected five (5) posters that represent emerging trends and important work in related technical areas. As usual, the conference features the latest processor designs for server and portable systems, multimedia and graphics, networking and telecommunications chips, and FPGA devices. The diverse program covers designs optimized for sub-threshold voltage operation all the way to designs exceeding 5GHz clock frequency. We are also happy to feature two excellent talks and four posters from academic projects.

Multi-core architectures and design for power efficiency remain the two most pervasive trends in the program. Nevertheless, specialization and heterogeneity are also emerging as important developments. Ten of the twenty-five talks in the program describe chips with multiple types of processing engines, programmable and fixed function. Heterogeneity is also the focus of the first tutorial that addresses the critical problem of software development for the heterogeneous multi-core chips in mobile devices. The second tutorial covers how die-stacking technology can improve latency, bandwidth, and system size, while preserving the benefits of heterogeneous manufacturing processes. Another exciting development this year is the appearance of chips that take established instruction sets beyond their traditional application domains. The program features talks on a smartphone chip based on the x86 ISA and a server chip based on ARM, in addition to talks on the latest designs based on the Power, SPARC, and MIPS instruction sets.

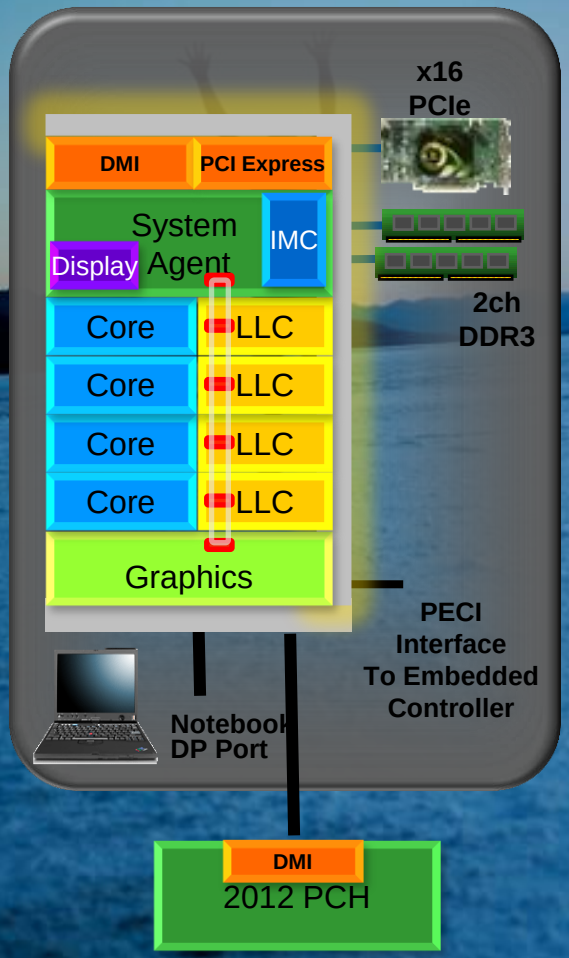
For the keynotes, we selected three exciting talks from leading figures in our industry. In the first keynote, Mark Papermaster will cover AMD's strategy towards heterogeneous systems and accelerated computing. In the second keynote, Marcus Weldon will discuss the future of wireless telecommunications and its implications to the semiconductor industry. The final keynote by Pat Gelsinger will discuss how cloud computing and big data are transforming the whole IT industry.

The high quality of this year's program is the direct result of the effort of the members of the program committee, all of whom worked hard to solicit, select, and improve presentations. We would also like to thank Liam Madden, Niel Trevett, Ralph Wittig, and Anwar Ghuloum for putting together the tutorials. The members of the organizing committee worked equally hard to provide the best possible setup for a successful symposium, overcoming several difficulties associated with the new location. An incredible amount of effort has gone into organizing tasks that we all take for granted such as high quality proceedings, online registration, and meals. Finally, we acknowledge the effort of all speakers, without whom there would be no conference.

Finally, we would like to recognize the contributions of Chuck Moore and John Nickolls that passed away recently. In addition to being leading visionaries and innovators in our field, Chuck and John were exemplary members of the Hot Chips community that contributed greatly through multiple roles. They will be missed.

Christos Kozyrakis and Rumi Zahir
Program Co-Chairs
Hot Chips 24
August 2012

Power Management of the Third Generation Intel Core Micro Architecture formerly codenamed Ivy Bridge



Sanjeev Jahagirdar
Varghese George, Inder Sodhi, Ryan Wells

Contents

- Ivy Bridge Overview
- Power Scaling & Efficiency
- Idle power Management
- Configurable TDP
- Clocking
- Additional Information

Contents

- Ivy Bridge Overview
- Power Scaling & Efficiency
- Idle power Management
- Configurable TDP
- Clocking
- Additional Information

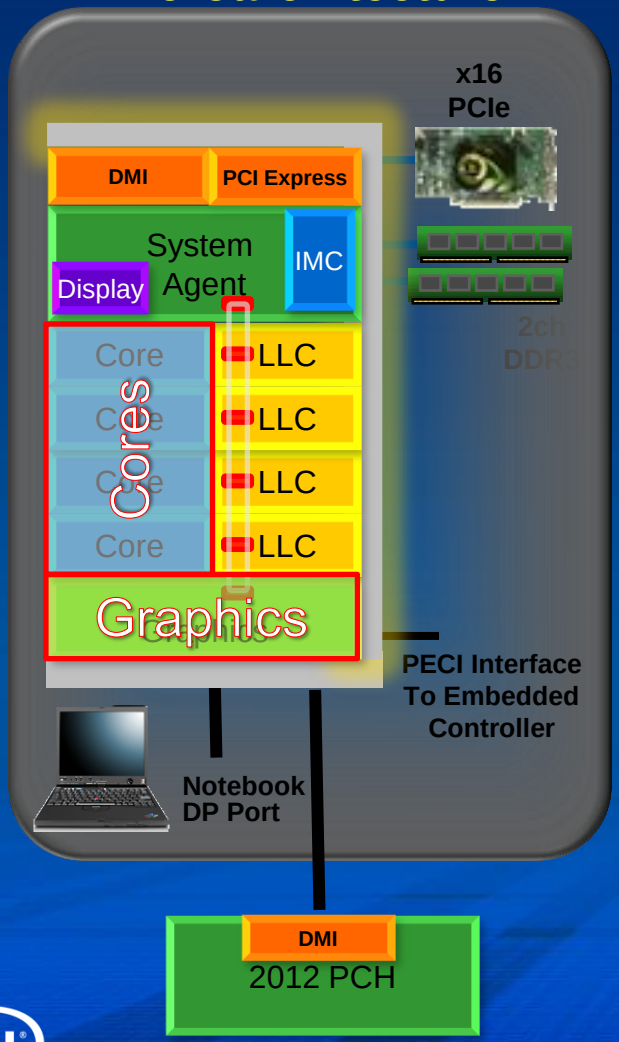
Intel's Tick-Tock Philosophy

- **Tock Processors**
 - Provide substantial microarchitecture improvement…
 - …on existing manufacturing process
- **Tick Processors**
 - Retain existing microarchitecture, …
 - …but utilize next generation fabrication technology to drive high volume and low product cost

- **The Tock: *Sandy Bridge***
 - Brought new ring/LLC microarchitecture
 - Integrated Graphics on ring
 - Integrated North Bridge ("System Agent"), including memory controller
- **The Tick: Ivy Bridge**
 - Process lead vehicle: Intel's 22nm process node
 - The Caveat:
 - Some Ivy Bridge areas have substantial (tock-like) change (Graphics)

Ivy Bridge – the 1st 22 nm Core Product

Ivy Bridge Microarchitecture



- **Leveraged from Sandy Bridge:**
 - Continue the 2-chip platform partition (CPU + PCH)
 - Fully integrated on silicon:
 - 2-4 IA Cores
 - Processor Graphics, Media, Display Engine
 - Integrated Memory Controller
 - PCIe Controllers
 - Modular On-Die Ring Interconnect
 - Shared LLC between IA Cores and Graphics
 - Same socket, similar packages
 - Similar SKUs (TDP, die configurations)
 - IVB backwards compatible with SNB

Ivy Bridge – Key New Things

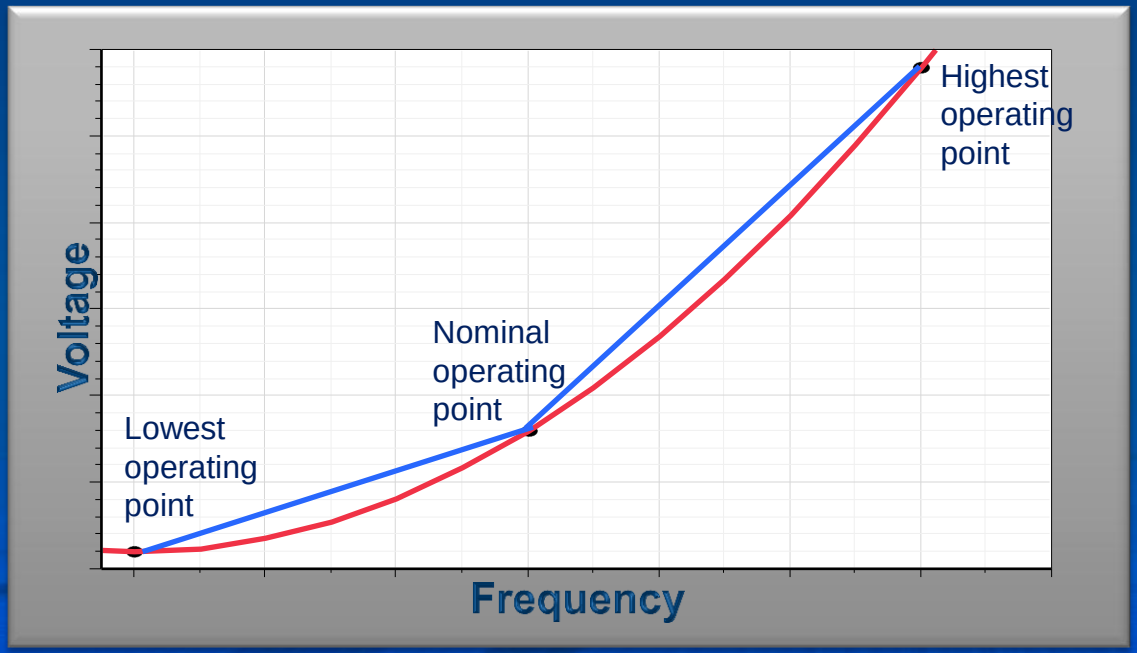
- **Entire chip moves to 22nm**
 - Higher performance/Lower power
- **Instruction Set Architecture Enhancements**
 - Float16 / Fast FS/GS support / REP MOVSB / RDRAND
- **Security Enhancements**
 - DRNG / SMEP
- **Power Improvements**
 - Scalability features: ConfigTDP
 - Average Power features: DDR power gates / PAIR
- **IO/Memory**
 - DDR3L support
 - Improved overclocking support
- **Performance Improvements (Instructions/clock)**

Contents

- Ivy Bridge Overview
- **Power Scaling & Efficiency**
- Idle power Management
- Configurable TDP
- Clocking
- Additional Information

Power efficiency via scaling & testing

- **Power Scaling in 22nm process extracted in two ways**
 - Higher performance in IA & Graphics within a power envelope
 - Lower operating Voltage in System Agent and Memory controller



- Power loss from discrete test points and interpolation (blue line)
- Ivy Bridge builds a quadratic model of the VF based on enhanced testing (red line)
- Optimal voltage at all operating points

Power efficiency via interrupt routing

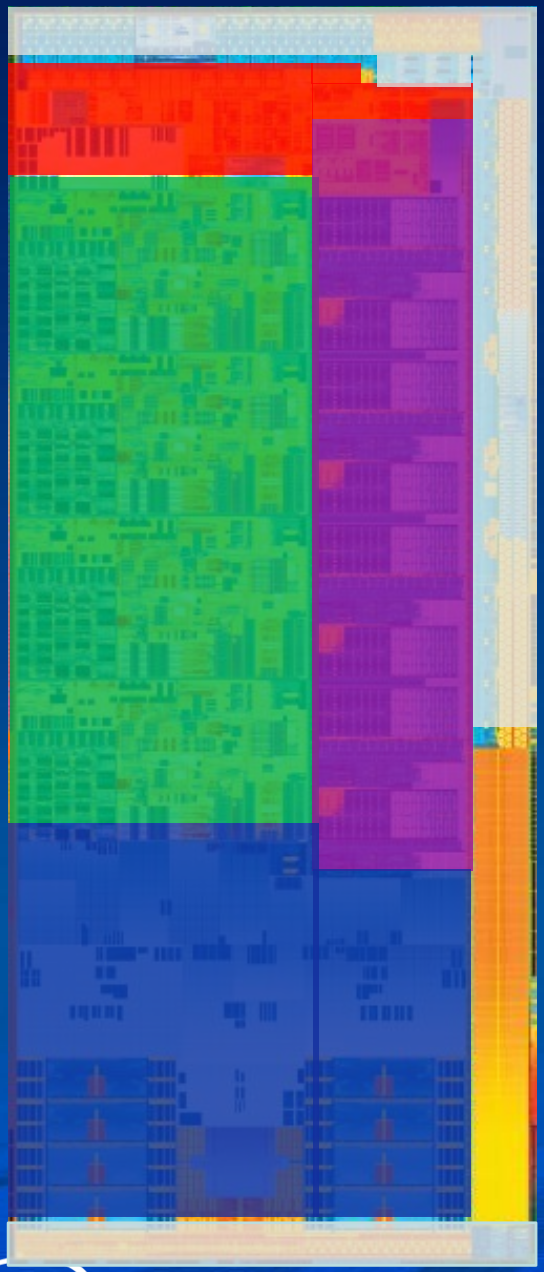
- **PAIR algorithm lowers power or performance impact of re-routable interrupts**
 - Compares power-state of all cores eligible to service interrupt
 - Chooses “best core” based on optimization mode (Power vs. Performance)
 - “Best Core” based on the following
 - Core C-states
 - P-state request (turbo vs. non-turbo)
- **Example: 1 core in C6 & 1 in C0**
 - Power bias will direct the interrupt to core in C0
 - Performance bias will wake the C6 core

Temperature effects

- **Thermal sensors are located in the hot spots in the IA core and GPU core**
- **Inverse temperature dependence (ITD) effects more pronounced in the 22nm node**
 - No sensors at the cold spots
 - IVB estimates the coldest point on the die to based on thermal sensors compensate for the effect
- **Manufacturing test voltages at hot and cold temperatures**
 - PCU interpolates linearly at run time to determine the voltage
 - Temperature moves slowly enough for the PCU and voltage regulator to keep up

Contents

- Ivy Bridge Overview
- Power Scaling & Efficiency
- **Idle power Management**
- Configurable TDP
- Clocking
- Additional Information



Ivy Bridge Power Planes

- **Key Power planes**
 - Core (Gated – Green)
 - LLC (Ungated – Purple)
 - SA/Display - Red
 - GT - Blue
 - Others (like IO, PLL etc) - Gray



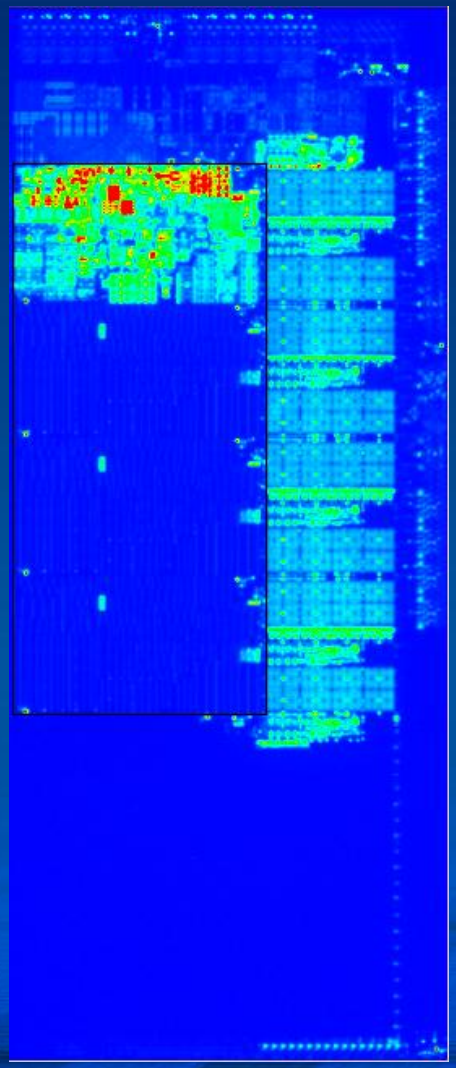
IVB Embedded Power Gate

Ivy Bridge has 3 on-die power gating areas

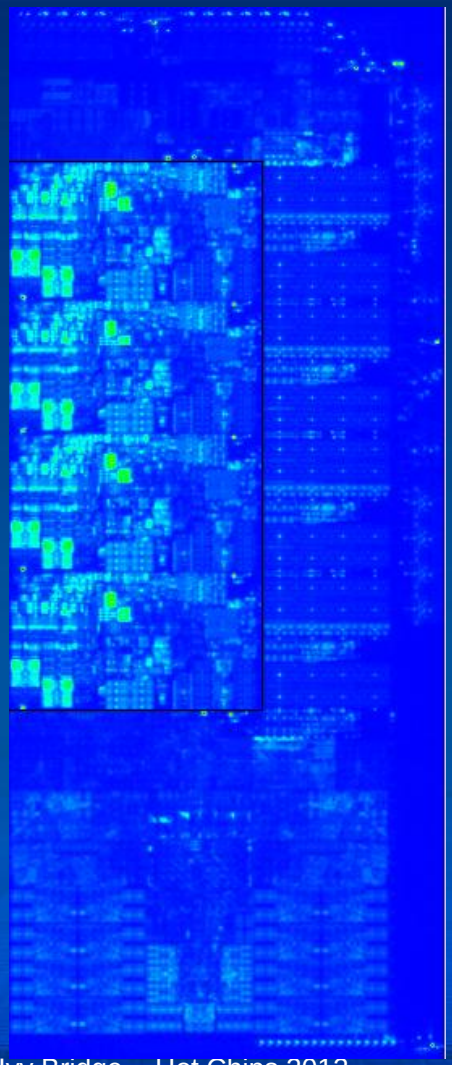
- Cores (Green)
 - Independent Gating per Core
 - Unified Cache
- PCIE controller(Red)
 - Gating static only when no connection
- DDR (Purple)
 - Gating of digital logic in the buffer applied during self-refresh mode

IREM images

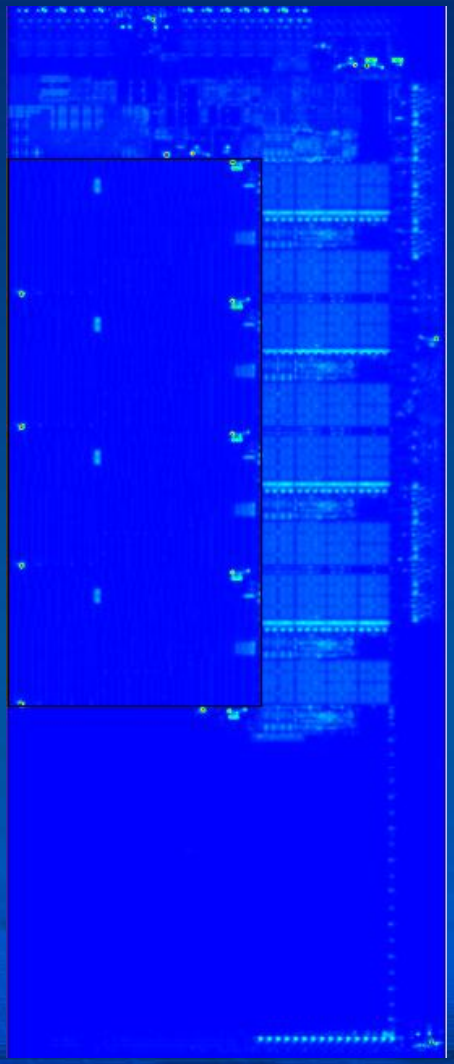
1 core in turbo, other 3 cores power gated



Typical Usage of Cores and Graphics



Cores and Graphics gated



DDR I/O Power Gating

- **Ivy Bridge implements on-die Embedded Power Gating (EPG) on DDR I/O**
- **Latency & Tradeoffs**
 - Latency considerations
 - Enabled on entry into Package C3 and deeper (memory in Self Refresh) to deal with latency of power gate
 - Additional latency of <5uS for device access to memory during exit
 - Conditional enabling – only if devices can tolerate the latency
 - No Impact to exit latency for interrupts
 - Design tradeoffs
 - To get around saving and restoring context, the DDR state is put on an ungated power island
 - For Idle/MM07-OP, Intel expects DDR IO to be gated ~90% of the time

Low Voltage optimizations

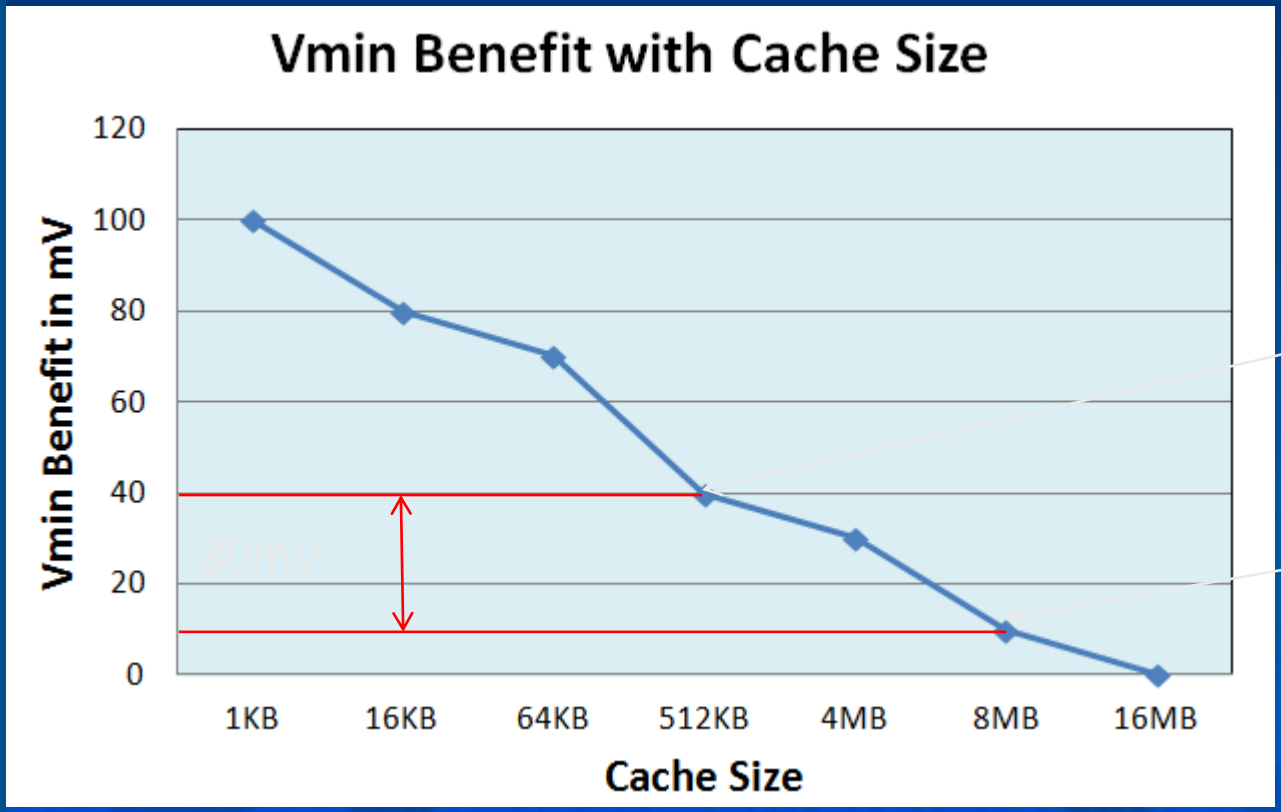
- **Small Signal arrays and register files limit the lowest operating voltage and retention voltage**

1. **Dynamic cache sizing to achieve a lower cache Vmin**
 - Cache Vmin is limited by ‘bad cells’ or defects distributed across the cache
 - A smaller size cache has a lower Vmin due to fewer defects
2. **PCU Firmware based register file re-initialization on exit from standby states**
 - Allows reduction of retention voltage below the retention level of the register file



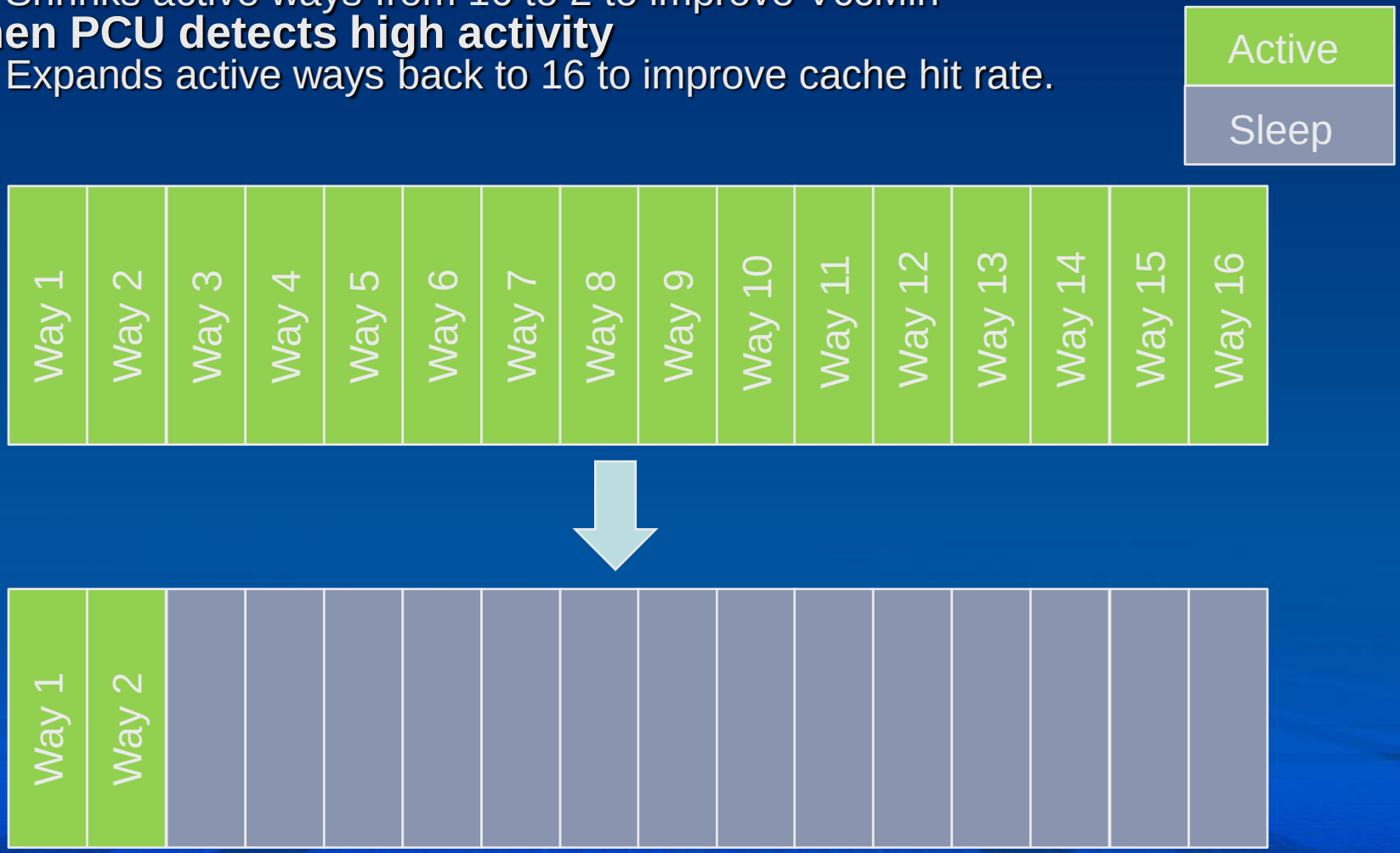
LLC - Dynamic Cache Shrink Feature

- Reduce LLC cache size dynamically from 8MB to 512KB to gain 30mV Vmin benefit
- LLC Expand/Shrink algorithm is developed for this purpose
- Entry/exit points were defined based on the work loads & performance

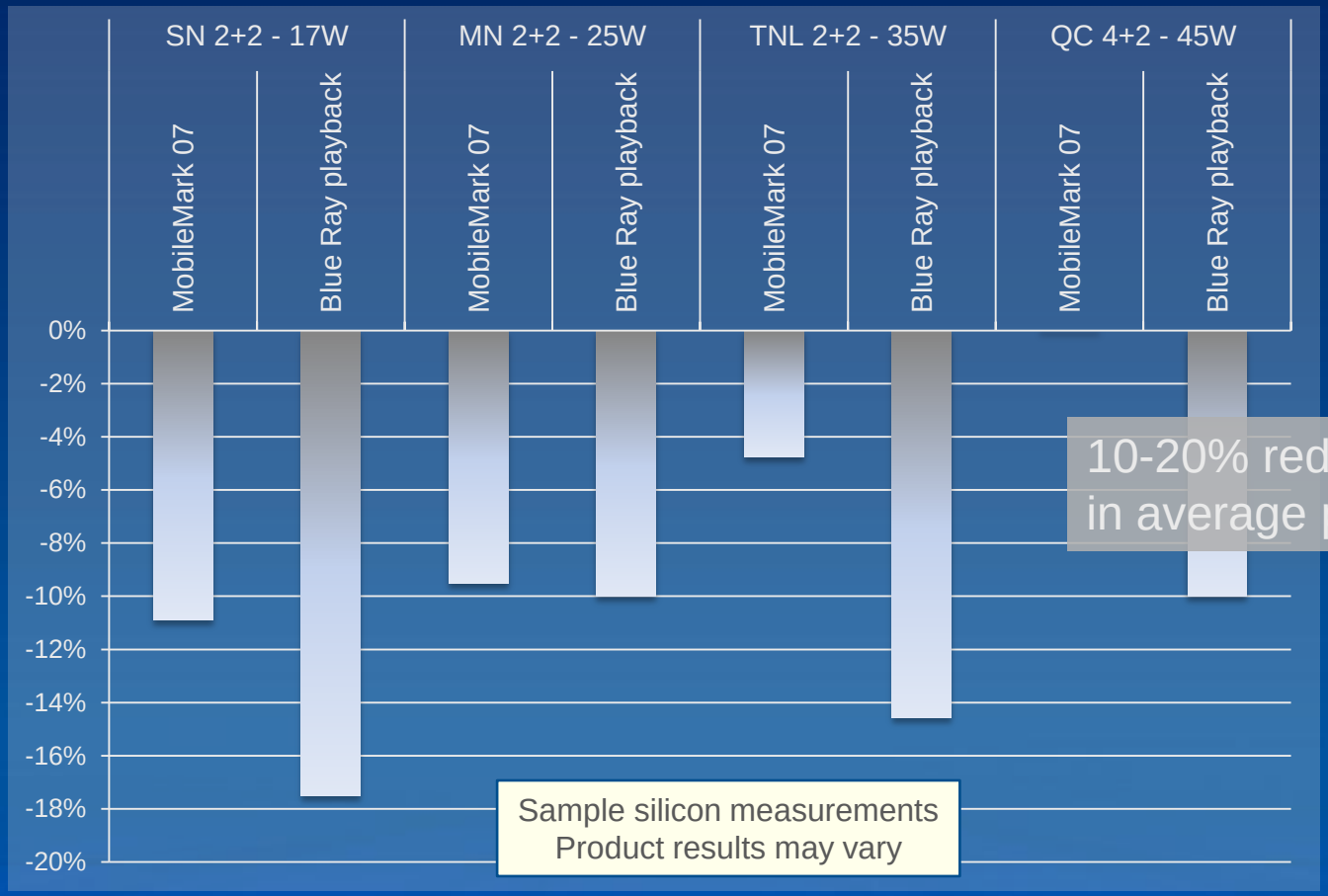


LLC - Dynamic Cache Shrink Feature

- **LLC organized in 16 ways.**
- **When PCU detects low activity workload**
 - Flushes 14 ways of the cache and puts ways to sleep
 - Shrinks active ways from 16 to 2 to improve VccMin
- **When PCU detects high activity**
 - Expands active ways back to 16 to improve cache hit rate.



Ivy Bridge average power reduction
(relative to SNB)



Power reduction via new PM features and process scaling benefits

Benefits on other SKUs varies

Contents

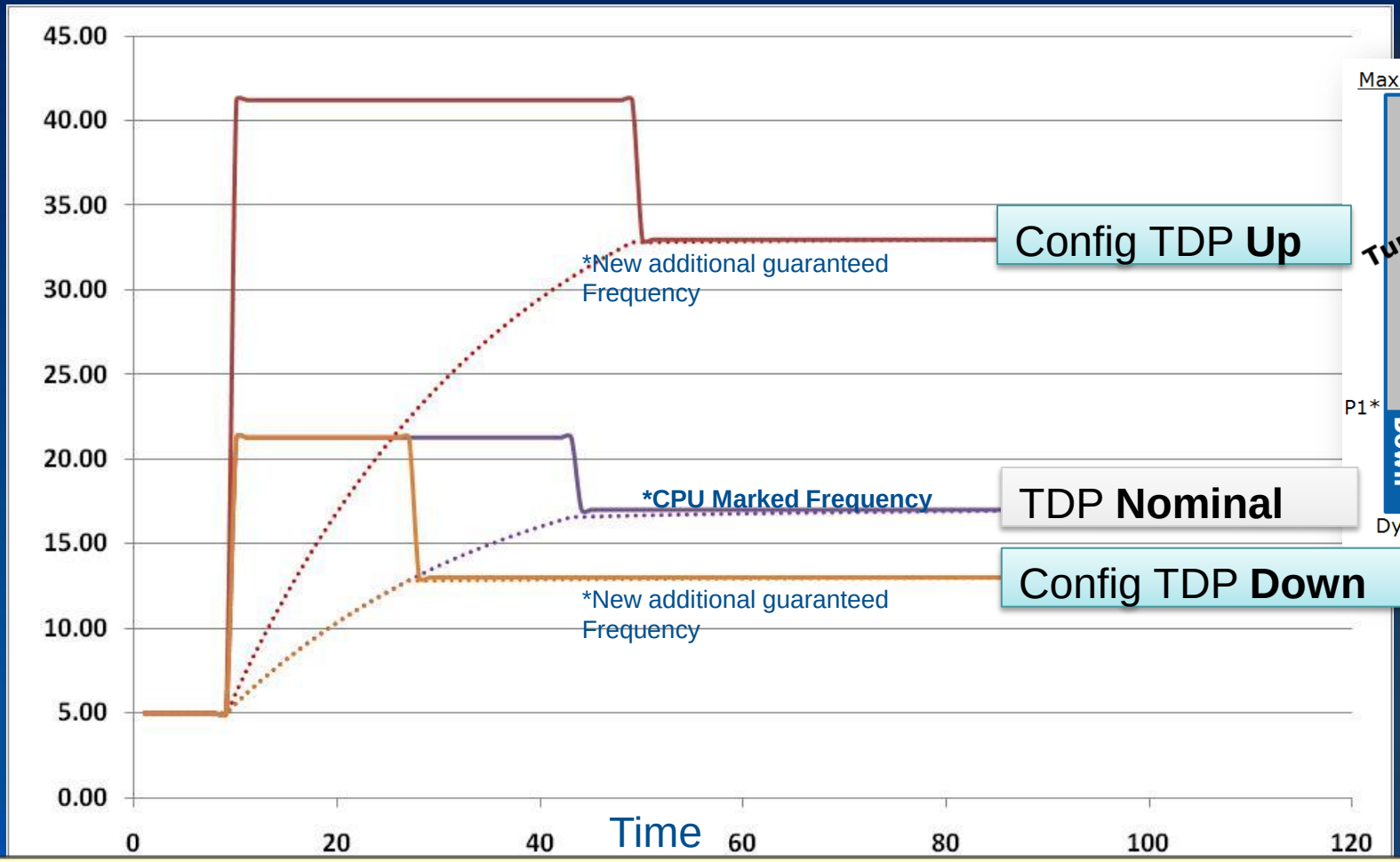
- Ivy Bridge Overview
- Power Scaling & Efficiency
- Idle power Management
- **Configurable TDP**
- Clocking
- Additional Information

Configurable TDP & Low Power Mode

- **Configurable TDP allows multiple TDP levels within the same part**
 - Greater dynamic range of power/performance guaranteed by Intel
 - Dynamically transition based on runtime triggers
- **Low Power Mode defines lowest active operating point for the part**
- **Intel offers software driver implementing both features**
 - System designers can utilize this framework and customize to their needs
- **Allow OEMs and End Users to take advantage of scalability of Intel CPUs**



cTDP Power Control

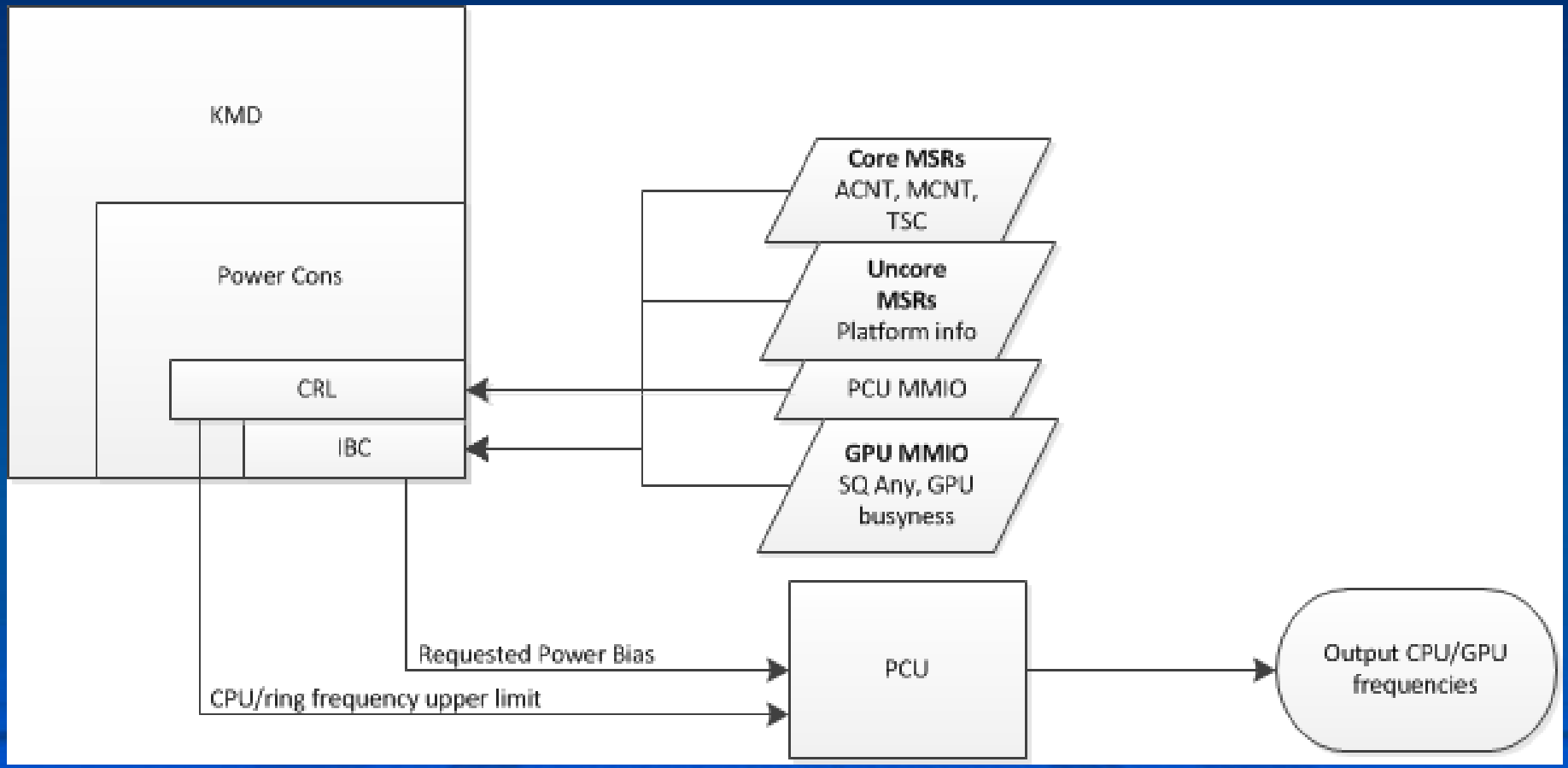


Regulated power limit adjusted in conjunction with TDP to allow guaranteed frequency (performance) at a specific power level

IA/GPU Power sharing

- **OEMs can configure the cooling limits to <17W**
 - Static biasing (X% to GPU and 100-X% to IA) results in sub optimal performance
- **Solution: Distribute power based on workload demand**
 - Determine target CPU/ring frequency based on workload
 - If actual CPU/ring freq < (target frequency – guard band)
 - *Move bias toward CPU Else Move bias toward GPU*
 - With hysteresis

Intelligent Bias Control Architecture



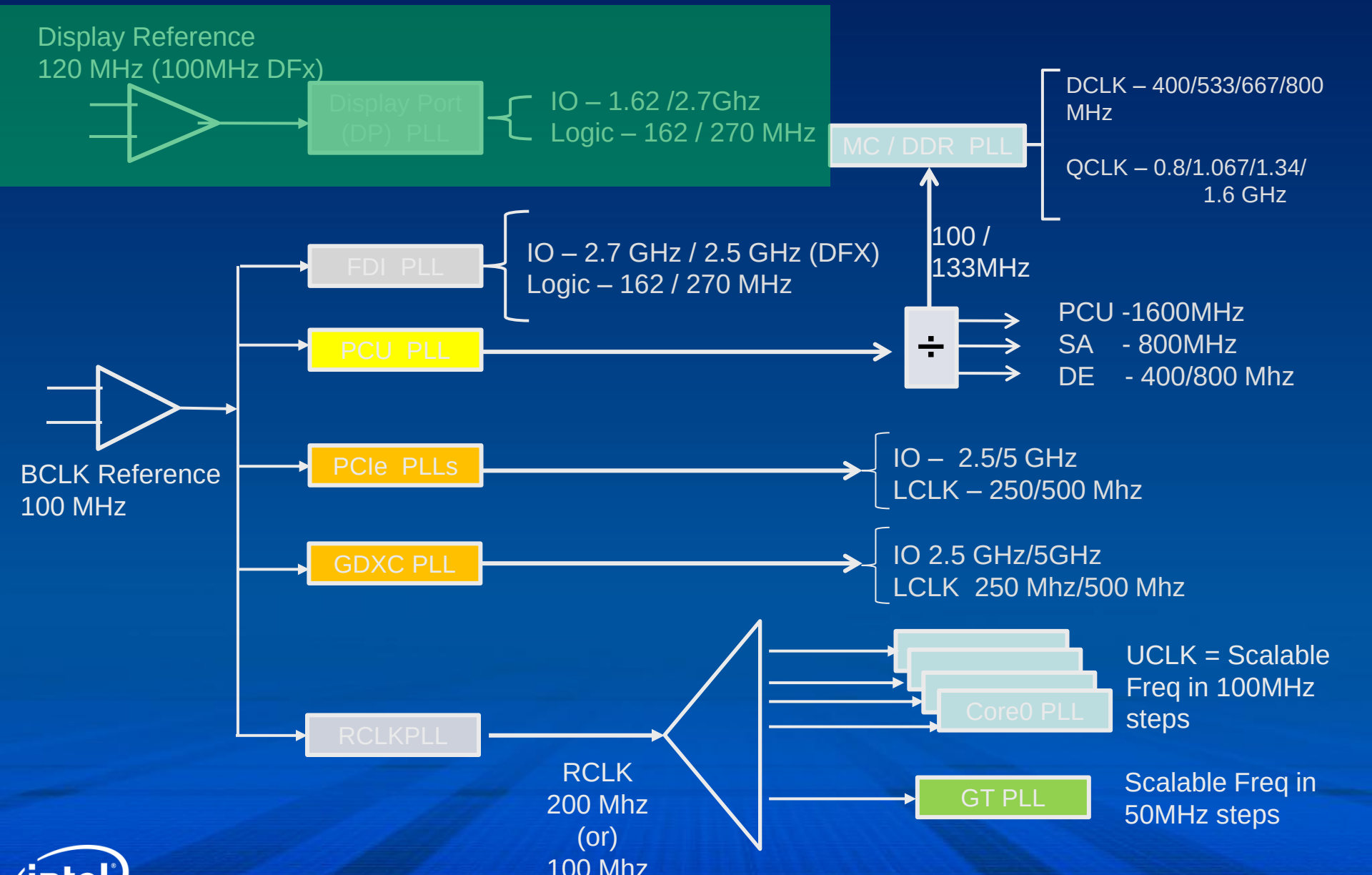
Platform Power management

- **Power delivery management**
 - How do we deal with the platform need to divert current from the CPU to other components dynamically?
 - IVB PCU will manage the current draw and will honor dynamic max current updates
- **Platform debug and tuning hooks**
 - IVB provides feedback to platform designers if power delivery, & cooling is limiting performance

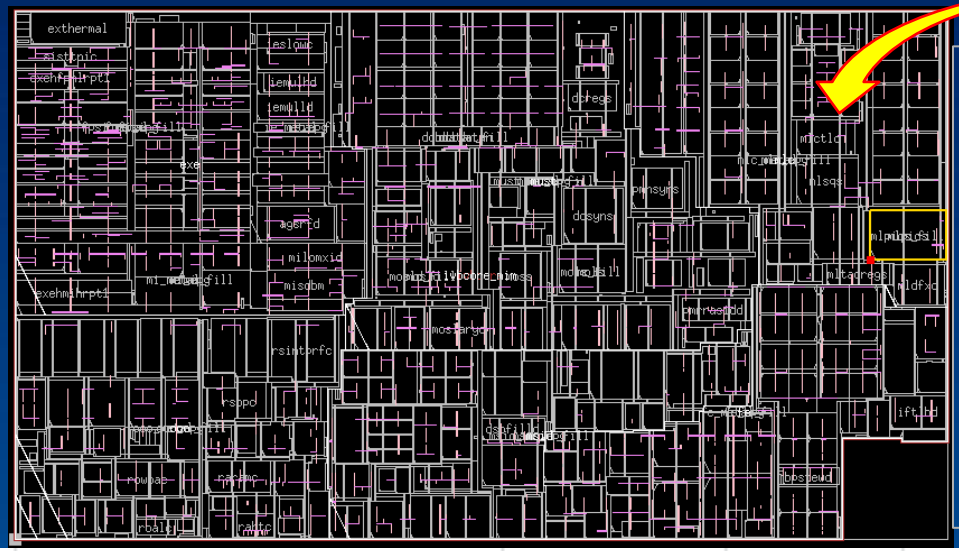
Contents

- Ivy Bridge Overview
- Power Scaling & Efficiency
- Idle power Management
- Configurable TDP
- **Clocking**
- Additional Information

IVB Clock Domains



PLL/Clocking



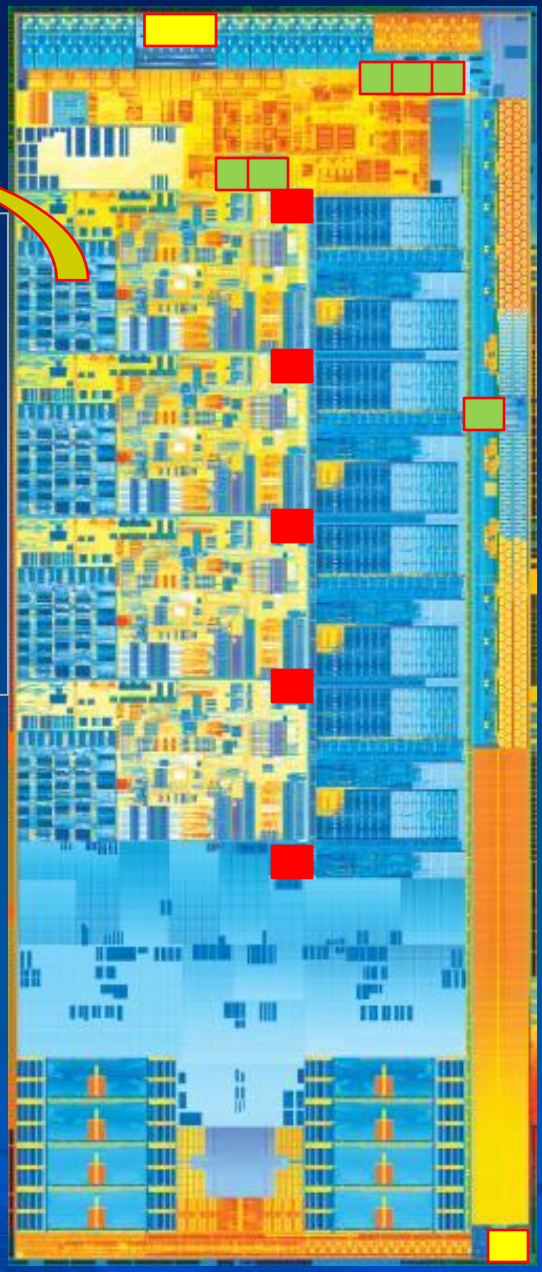
Clock Islands in Core

Each Island can be independently clock gated.

Clock Islands in Core =180

Clock Islands in LLC = 48

Slice Clocking Cdyn	CORE (pF)	L3(pF)	TOTAL(pF)
Total (including RT)	179	77	256
Global Drivers + Islands	109	46	155
Clock Source + Spines	70	31	101
Silicon Measurement	181	81	262

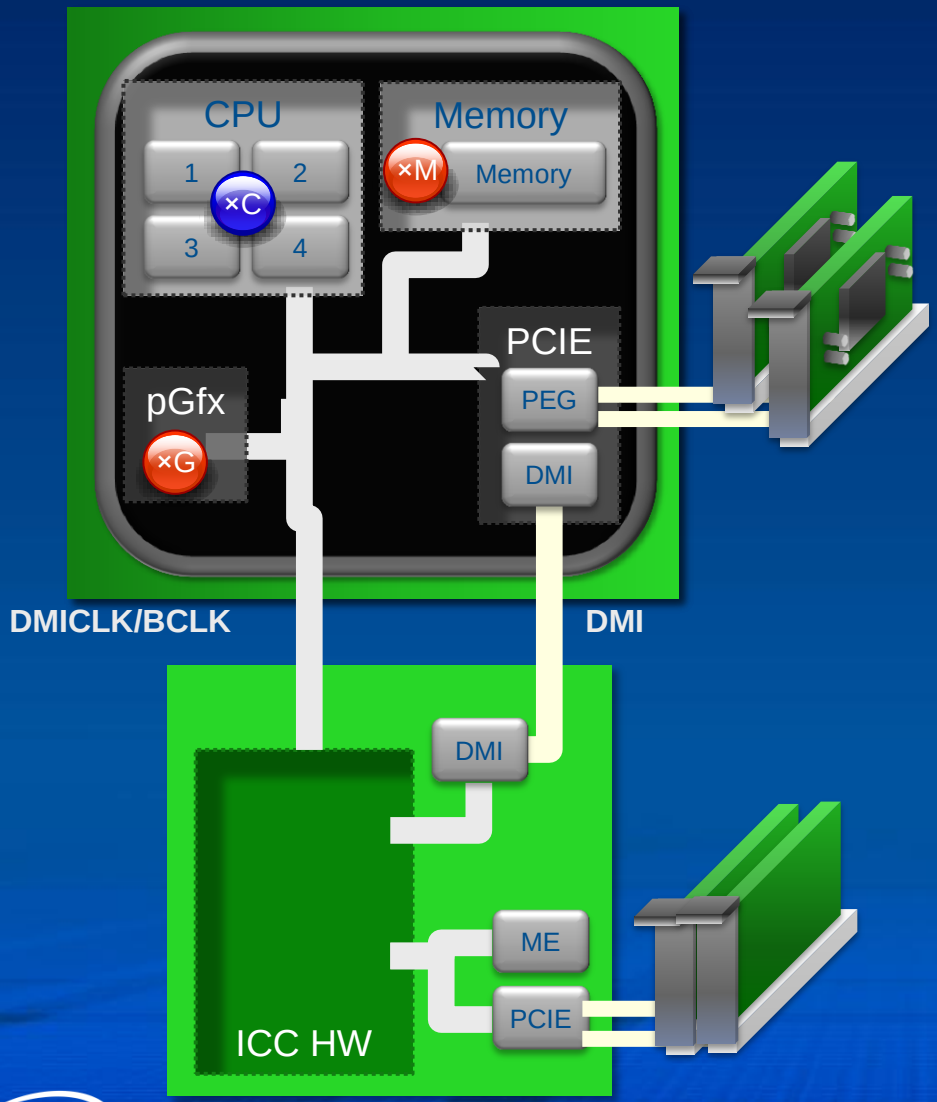


Wide Range SB PLL

PCIE LC PLL

Single Ratio SB PLL

Overclocking Enhancements

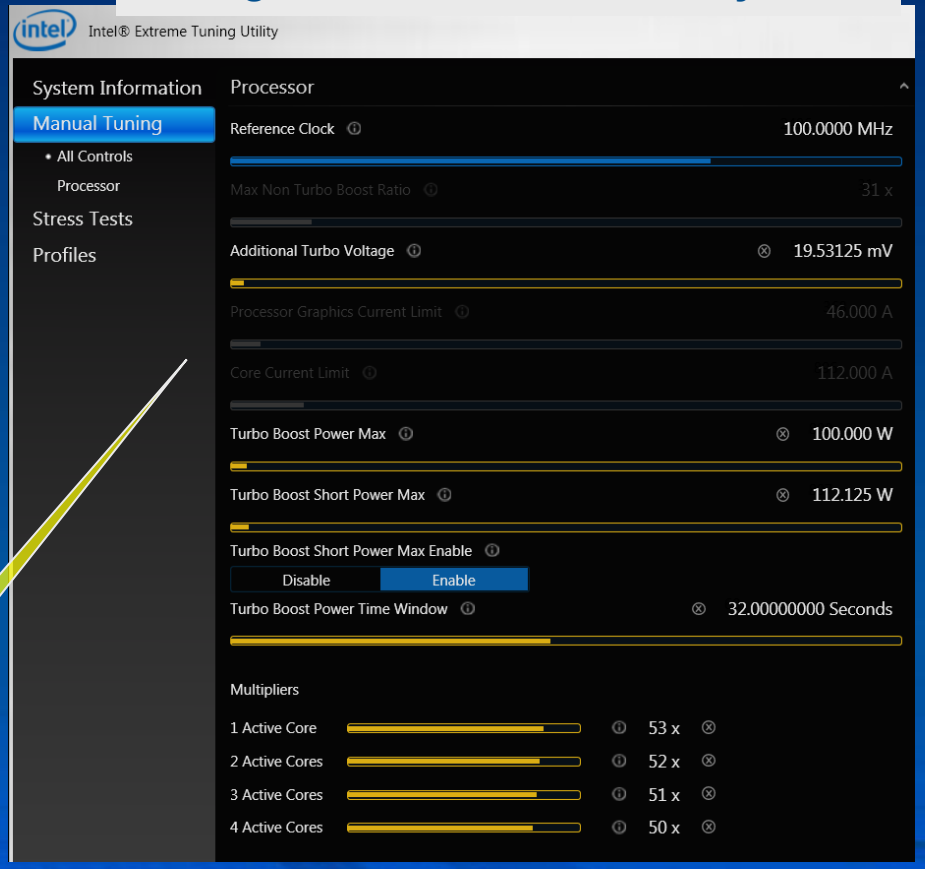


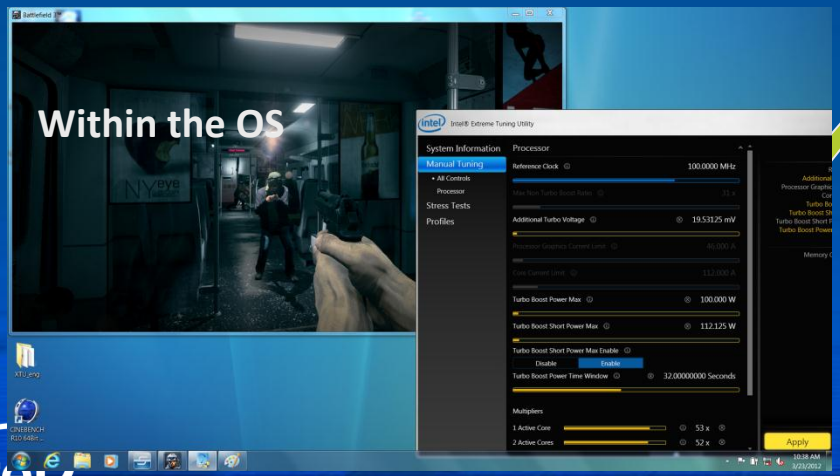
- Core Frequency ×C
 - Unlocked turbo limits
 - Unlocked core ratios up to 63 in 100MHz increments†
 - Programmable voltage offset
- Graphics Frequency ×G
 - Unlocked graphics turbo limits
 - Unlocked graphics ratios up to 60 in 50MHz increments
 - Programmable voltage offset
- Memory Ratio ×M
 - Unlocked memory controller
 - Granularity options for 200 and 266MHz
 - Logical support up to 2666MHz
- DMICLK (aka BCLK)
 - Unlocked PCH clock controller (1MHz increments)
- PEG and DMI
 - Fixed ratios

Real-Time Overclocking

- PCU samples OC parameters continuously and updates power limits
- OC without reboot:
 - Maximum Core Ratio
 - Processor Graphics Ratio
 - BCLK (small increments)
 - Power Limits: PL1, PL2, Tau
 - Additional Turbo Voltage for CPU and pGfx

Changes effective immediately





Acknowledgements

- Authors would like to thank the entire Ivy Bridge team for their dedicated work.

Contents

- Ivy Bridge Overview
- Power Scaling & Efficiency
- Idle power Management
- Configurable TDP
- Clocking
- Additional Information

Ivy Bridge ISA & Security enhancements

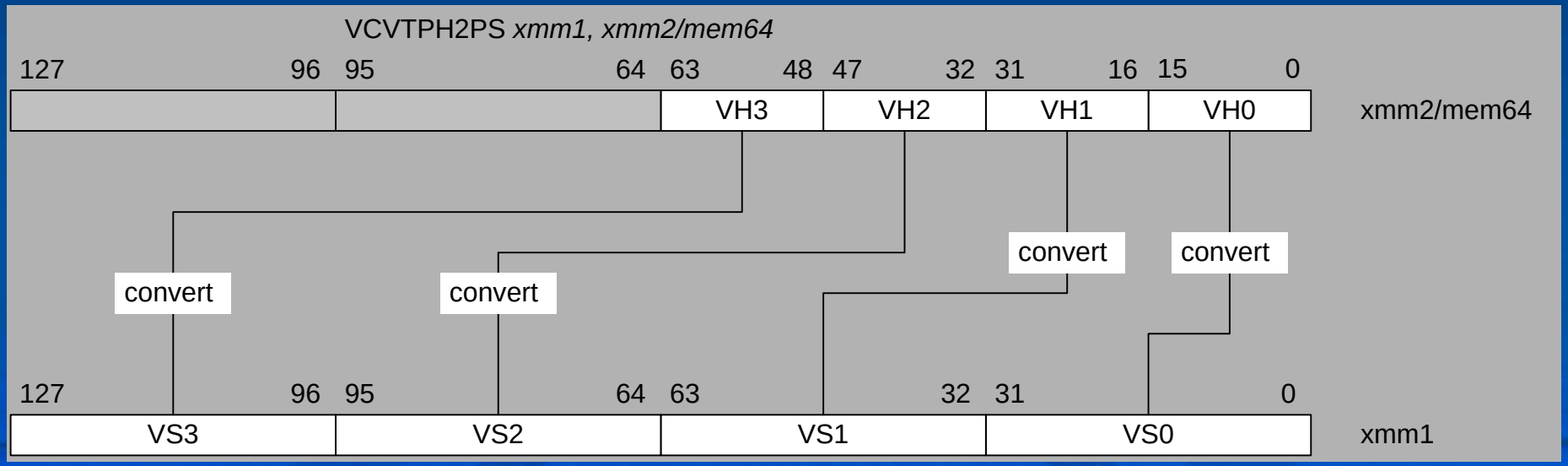
Float16 Data Conversion Instructions

- **New instructions for supporting conversion between a 16-bit floating point memory format and 32-bit single precision**
 - VCVTPH2PS, VCVTPS2PH
 - Both 128 (SSE) and 256 bit (AVX) wide vector flavors supported
 - Only supported in the VEX prefix context
- **Facilitates use of single-precision floating point computations from a more compressed memory format**
 - 1-bit sign, 5-bit exponent, 10-bit significand (+ implicit integer bit)
- **Enables higher dynamic range compared to fixed point within the same storage footprint**
 - Image processing, video decode, audio processing
 - 50% reduction in storage v. single-precision FP (w/ loss of fidelity)
- **Enumerated via new CPUID feature flag**
 - CPUID.1.ECX[29]

VCVTPH2PS – Convert 16-bit float to SP

VCVTPH2PS *ymm1, xmm2/mem128* - *256 bit vector*
VCVTPH2PS *xmm1, xmm2/mem64* - *128 bit vector*

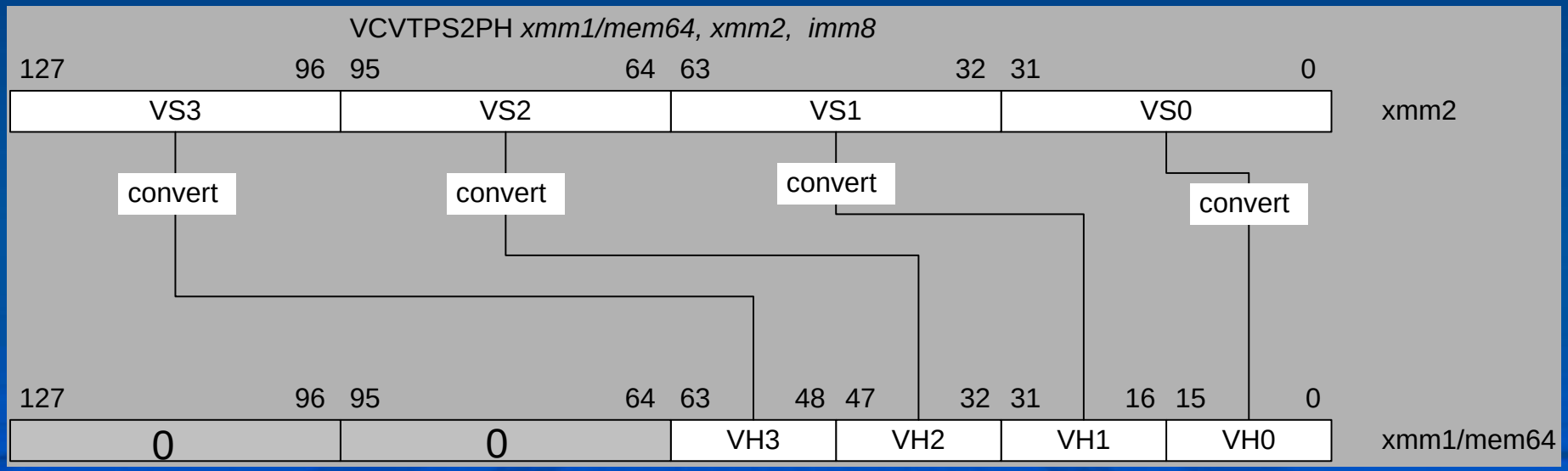
Converts four packed 16-bit floating-point values in the low 64 bits of XMM2 or 64-bit memory location to four single-precision floating-point values and writes the results in the destination (XMM1 register).



VCVTPS2PH – Convert SP to 16-bit float

VCVTPS2PH *xmm1/mem64, xmm2, imm8* - *128 bit vector*
VCVTPS2PH *xmm1/mem128, ymm2, imm8* - *256 bit vector*

Converts four packed single-precision floating-point values in XMM2 to four 16-bit floating-point values and writes the results in the destination (XMM1 register or memory location).



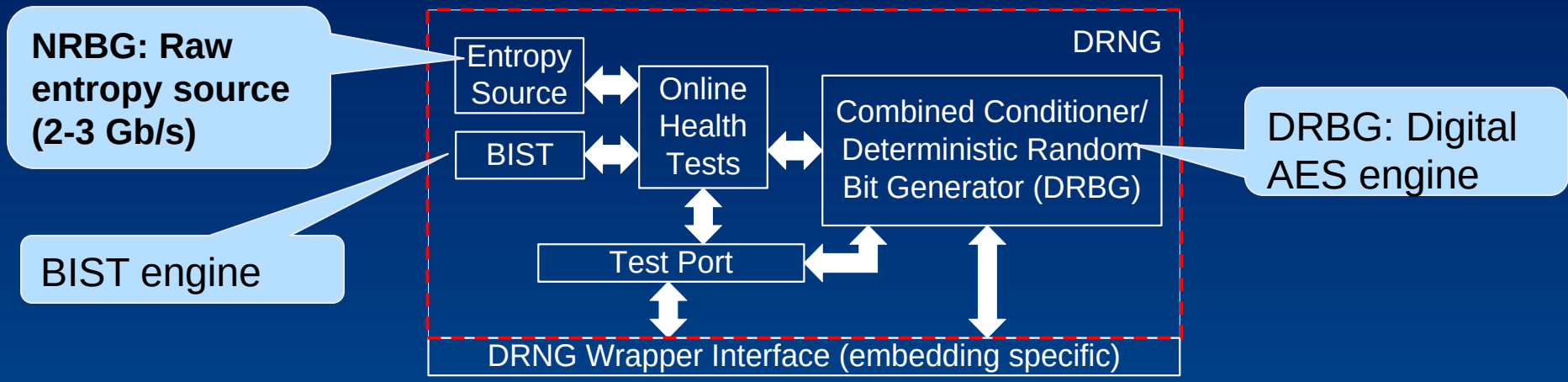
Write/Read FS/GS Base Instructions

- **New ring-3 instructions for read/write of the FS & GS segment base registers**
 - To be used by user level code for thread local storage
 - Enumerated via new CPUID feature flag
 - CPUID.7.0.EBX[0] indicates availability (leaf 7, subleaf 0)
 - Requires enabling by OS to permit FS/GS segment base access
 - CR4.RDWRGSFS (bit 16) = 0 (default)
- **Motivation:**
 - Improve scalability and programming ease for user threads

REP MOVSB/STOSB improvements

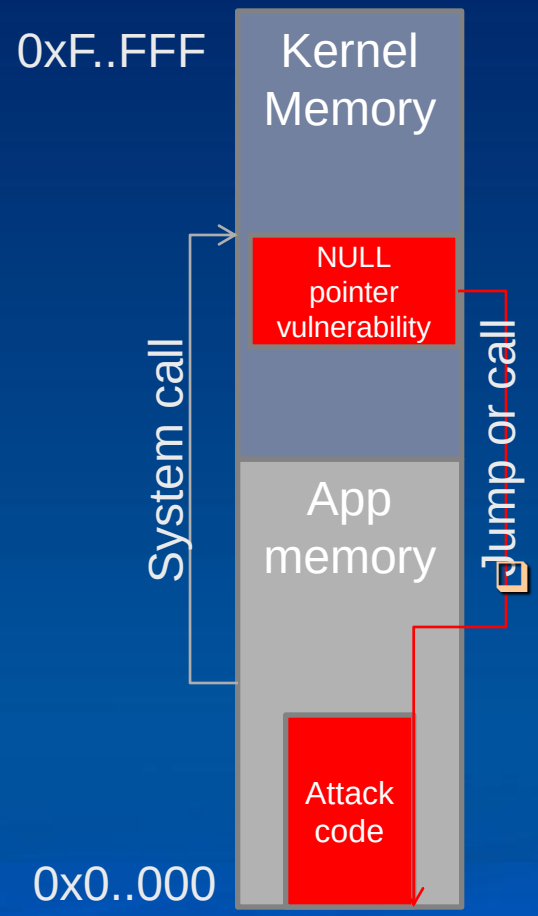
- **Historically optimizing block copy/fill operations tends to be microarchitecture specific**
 - Lack of a “one size fits all” solution implies CPU model specific algorithms for best performance
- **IVB address this through more optimized REP MOVSB and REP STOSB instructions**
 - Expect this to replace the need for manual tuning solutions
 - Limitation: If block size is known at compile time and size <=64 bytes, then scalar loads & stores are still considered faster
- **Enhancement availability indicated by CPUID.7.0.EBX[9] (ENFSTRG)**
 - This bit can be used by run time SW (Libraries, JIT) for tuning to a specific implementation

Digital Random Number Generator (DRNG)



- **Background:**
 - Entropy is valuable in a variety of uses – Example: “keying material” in cryptography
 - Historically, computing platforms did not have a good source of a high quality/high performance “entropy source”
 - Typical sources used today are slow (bit rate in Kb/s) (key strokes, mouse clicks etc)
- **IVB introduces high quality/high performance DRNG**
- **The DRNG is designed to be Standards compliant**
 - ANSI X9.82,NIST SP 800-90 and NIST FIPS 140-2/3 Level 2 certifiable entropy source
- **New instruction: RDRAND – Available at all privilege levels/operating modes**
 - Instruction will return a random number (16, 32 or 64-bit) to the destination register
- **New CPUID feature flag for RDRAND enumeration**
 - CPUID.1.ECX[30]

Supervisory Mode Execute Protection (SMEP)



- **Background:**
 - Privilege Escalation Attack causes CPL 0 access to user mode pages
 - Example:
 - Step 1: Compromise user mode app or trick user into installing attack app
 - Step 2: Exploit OS vulnerability to force control transfer to user mode attack code while CPU remains in supervisory mode => privilege escalation
- IVB introduces SMEP to help prevent such attacks
 - Prevents execution of user mode pages while in supervisor mode
 - If CR4.SMEP set to 1 and in supervisor mode (CPL<3), instructions may not be executed from a linear address for which the user mode flag is 1
 - Available in both 32- and 64-bit operating modes
 - SMEP is enumerated via CPUID.7.0.EBX[7]

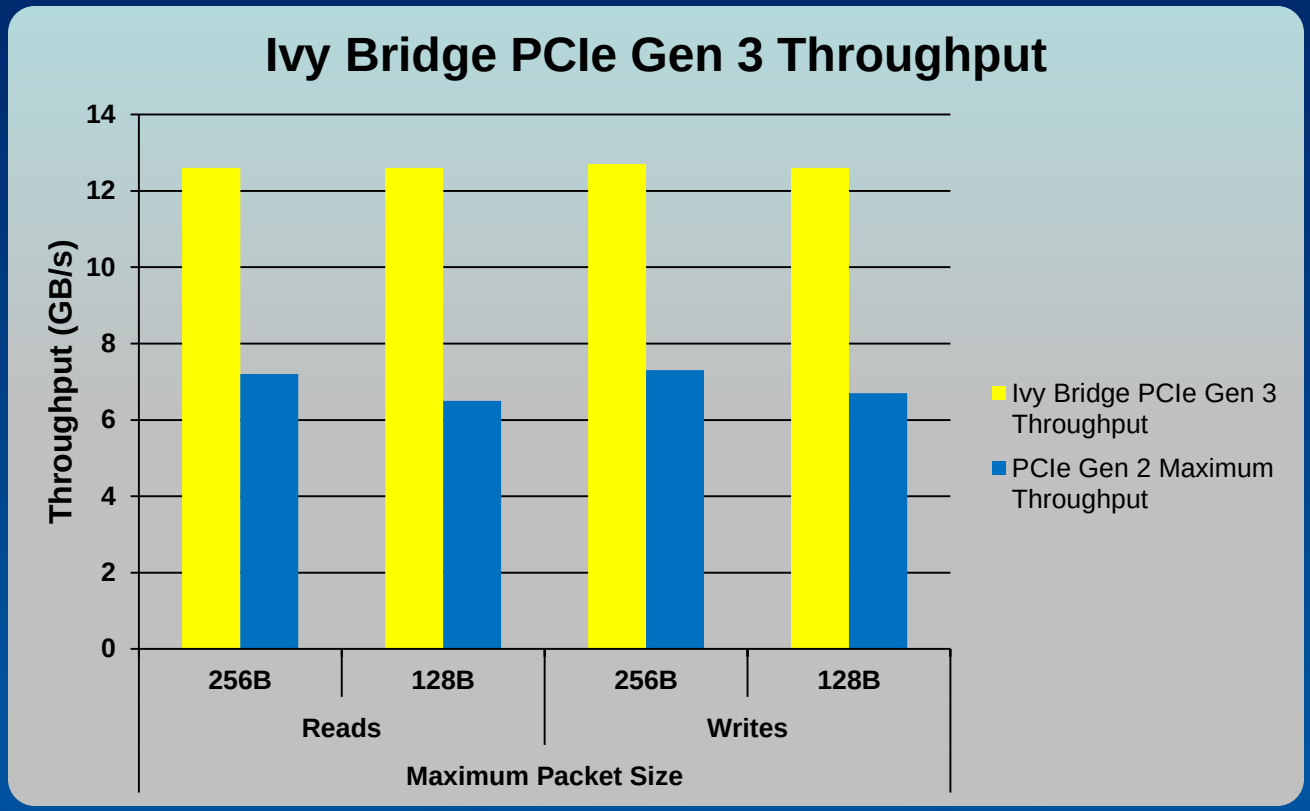
39

PCI Express Gen 3

Ivy Bridge PCI Express Gen 3

- **Third generation of the PCI Express I/O interface**
 - Delivers nearly twice the I/O bandwidth v. Gen 2
 - Improves performance for applications sensitive to I/O bandwidth
 - Enables smaller form factors via narrower, faster physical links
- **Bandwidth realized through:**
 - Faster signaling speed: 8 GT/s
 - More efficient lane encoding: 128/130
- **Utilizes Gen 2 I/O channel characteristics**
 - Enables compatibility with previous Gen components
 - Enables drop-in upgrade for Sandy Bridge-based platforms
- **Supports PCIe bandwidth management & ASPM states**
 - Dynamic Link Width Configuration, L0s (Rx & Tx), L1

Ivy Bridge PCIe Performance*



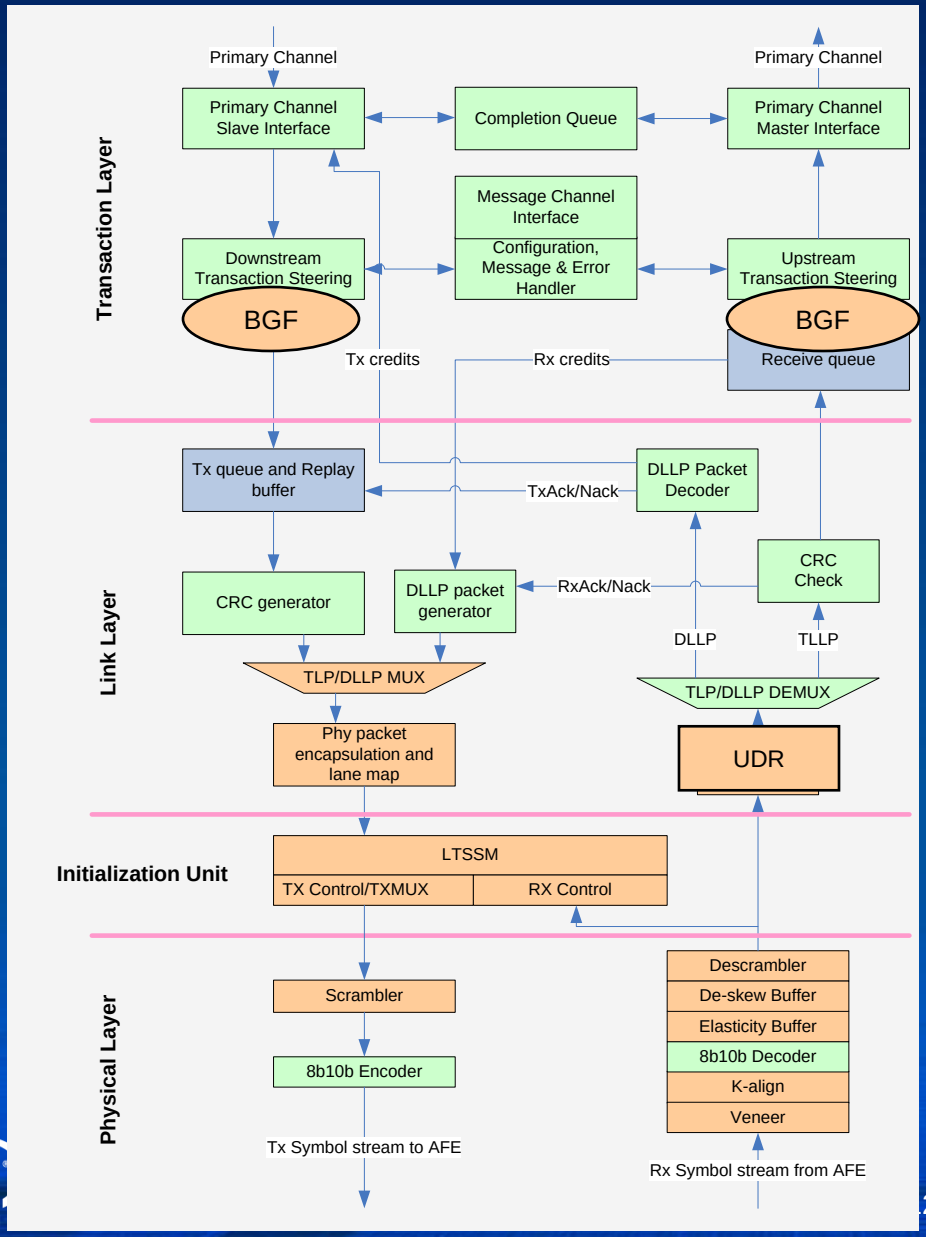
- **Ivy Bridge delivers nearly 2x Gen 2 bandwidth**
 - At similar latencies
 - ~300ns typical for upstream read request

Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

Ivy Bridge PCIe Logic Changes

800 MHz

500MHz/1GHz



- Sandy Bridge PCIe uArch unchanged
 - No change to primary channel hub/TL interface
 - No change to controller/PHY lanes interface
- Gen 3 changes layered on top of Gen 2 functionality (additional states, arcs)
 - Parallel flows implemented where feasible

No change

Logic change

Buffer size change

* Elasticity and De-skew buffers have both logic and size change

IPC Improvements

Most Significant IVB IPC Improvements

- **Pipeline MOV elimination**
 - Eliminates Move related micro-operations from the processor execution pipeline
- **Pipelined divider**
 - Improves throughput of divide related computations
- **Next page prefetcher**
 - Enables prefetching to span across a 4K page boundary
- **Shift/Rotate performance**
 - Addresses glass jaw concern with crypto and hashing algorithms
 - Addresses clumsiness of partial flag handling
- **6 additional split load registers**
 - Improves performance for loads splitting cache lines
 - Especially critical for AVX or SSE

Uncore IPC Features

- **AFP – Adaptive Fill Policy**
 - Cache heuristics to identify and segregate streaming applications
- **QLRU – Quad-Age LRU algorithm**
 - Allows fine-grain “age assignment” on cache allocation
 - E.g.: prefetched requests are allocated at “middle age”
- **DPT – Dynamic Prefetch Throttling**
 - Real-time memory bandwidth monitor
 - Directs core prefetchers to reduce prefetch aggressiveness during high memory load scenarios
- **Channel Hashing -- DRAM channel selection mechanism**
 - Allows channel selection to be made based on multiple address bits
 - Historically, it had been “A[6]”
 - Allows more even distribution of memory accesses across channels

Legal Notices and Disclaimers

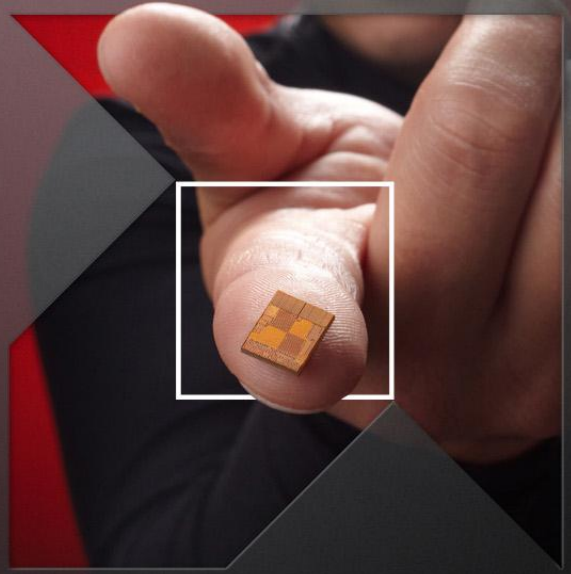
- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.
- Intel may make changes to specifications and product descriptions at any time, without notice.
- All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.
- Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.
- Any code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel's internal code names is at the sole risk of the user.
- Intel product plans in this presentation do not constitute Intel plan of record product roadmaps. Please contact your Intel representative to obtain Intel's current plan of record product roadmaps.
- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to http://www.intel.com/performance
- Intel, Intel Inside, the Intel logo, Centrino, Intel Core, Intel Atom, Pentium and UltraBook are trademarks of Intel Corporation in the United States and other countries.
- **Material in this presentation is intended as product positioning and *not* approved end user messaging.**
- **This document contains information on products in the design phase of development.**
- *Other names and brands may be claimed as the property of others.
- Copyright © 2012 Intel Corporation, All Rights Reserved

Legal Notices and Disclaimers, cont.

- WiMAX connectivity requires a WiMAX enabled device and subscription to a WiMAX broadband service. WiMAX connectivity may require you to purchase additional software or hardware at extra cost. Availability of WiMAX is limited, check with your service provider for details on availability and network limitations. Broadband performance and results may vary due to environment factors and other variables. See www.intel.com/go/wimax for more information.
- Intel® My WiFi Technology is an optional feature and requires additional software and a Centrino® wireless adapter. Wi-Fi devices must be certified by the Wi-Fi Alliance for 802.11b/g/a in order to connect. See mywifi.intel.com for more details.
- Hyper-Threading Technology requires a computer system with a processor supporting HT Technology and an HT Technology-enabled chipset, BIOS and operating system. Performance will vary depending on the specific hardware and software you use. For more information including details on which processors support HT Technology, see here
- Intel® Turbo Boost Technology requires a PC with a processor with Intel Turbo Boost Technology capability. Intel Turbo Boost Technology performance varies depending on hardware, software and overall system configuration. Check with your PC manufacturer on whether your system delivers Intel Turbo Boost Technology. For more information, see http://www.intel.com/technology/turboboost
- Requires an Intel® Wireless Display enabled PC, TV Adapter, and compatible television. Available on select Intel® Core processors. Does not support Blu-Ray or other protected content playback. Consult your PC manufacturer. For more information, see www.intel.com/go/wirelessdisplay
- (Built-in Visuals) Available on the 2nd gen Intel® Core™ processor family. Includes Intel® HD Graphics, Intel® Quick Sync Video, Intel® Clear Video HD Technology, Intel® InTru™ 3D Technology, and Intel® Advanced Vector Extensions. Also optionally includes Intel® Wireless Display depending on whether enabled on a given system or not. Whether you will receive the benefits of built-in visuals depends upon the particular design of the PC you choose. Consult your PC manufacturer whether built-in visuals are enabled on your system. Learn more about built-in visuals at http://www.intel.com/technology/visualtechnology/index.htm.
- Intel® Insider™ is a hardware-based content protection mechanism. Requires a 2nd generation Intel® Core™ processor-based PC with built-in visuals enabled, an Internet connection, and content purchase or rental from qualified providers. Consult your PC manufacturer. For more information, visit www.intel.com/go/intelinsider.
- Viewing Stereo 3D content requires 3D glasses and a 3D capable display. Physical risk factors may be present when viewing 3D material

Legal Notices and Disclaimers, cont.

- Security features enabled by Intel® AMT require an enabled chipset, network hardware and software and a corporate network connection. Intel AMT may not be available or certain capabilities may be limited over a host OS-based VPN or when connecting wirelessly, on battery power, sleeping, hibernating or powered off. Setup requires configuration and may require scripting with the management console or further integration into existing security frameworks, and modifications or implementation of new business processes. For more information, see http://www.intel.com/technology/manage/iamt.
- No system can provide absolute security under all conditions. Requires an enabled chipset, BIOS, firmware and software and a subscription with a capable Service Provider. Consult your system manufacturer and Service Provider for availability and functionality. Intel assumes no liability for lost or stolen data and/or systems or any other damages resulting thereof. For more information, visit http://www.intel.com/go/anti-theft
- Requires an Execute Disable Bit enabled system. Check with your PC manufacturer to determine whether your system delivers this functionality. For more information, visit http://www.intel.com/technology/xdbit/index.htm
- Intel® vPro™ Technology is sophisticated and requires setup and activation. Availability of features and results will depend upon the setup and configuration of your hardware, software and IT environment. To learn more visit: http://www.intel.com/technology/vpro
- The original equipment manufacturer must provide TPM functionality, which requires a TPM-supported BIOS. TPM functionality must be initialized and may not be available in all countries.
- Intel® AES-NI requires a computer system with an AES-NI enabled processor, as well as non-Intel software to execute the instructions in the correct sequence. AES-NI is available on select Intel® processors. For availability, consult your reseller or system manufacturer. For more information, see http://software.intel.com/en-us/articles/intel-advanced-encryption-standard-instructions-aes-ni/
- No system can provide absolute security under all conditions. Requires an Intel IPT enabled system, including a 2nd generation Intel Core processor, enabled chipset, firmware, and software. Available only on participating websites. Consult your system manufacturer. Intel assumes no liability for lost or stolen data and/or systems or any resulting damages. For more information, visit http://www.ipt.intel.com



"JAGUAR"

AMD's Next Generation Low Power x86 Core

Jeff Rupley, AMD Fellow
Chief Architect / Jaguar Core
August 28, 2012

AMD

TWO X86 CORES TUNED FOR TARGET MARKETS

“Bulldozer Family”
Performance & Scalability

“Cat Family”
Flexible, Low Power & Small

Mainstream Client and Server Markets






Small Die Area



Low Power Markets



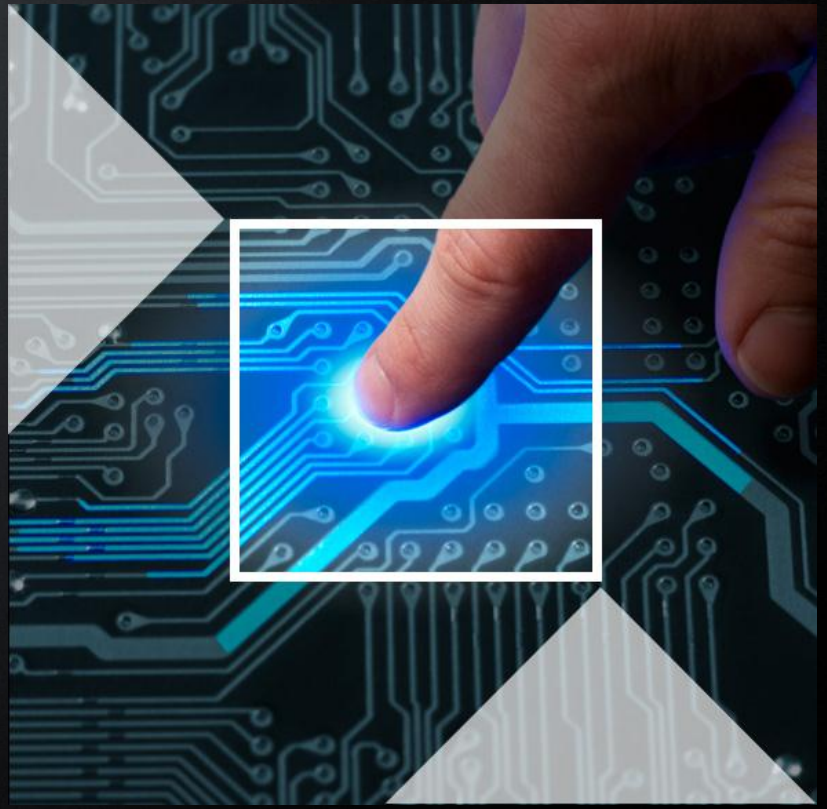


Cloud Clients Optimized

AMD

"JAGUAR" DESIGN GOALS

- Improve on “Bobcat”: performance in a given power envelope
 - More IPC
 - Better Frequency at given Voltage
 - Improve power efficiency thru clock gating and unit redesign
- Update the ISA/Feature Set
- Increase Process Portability

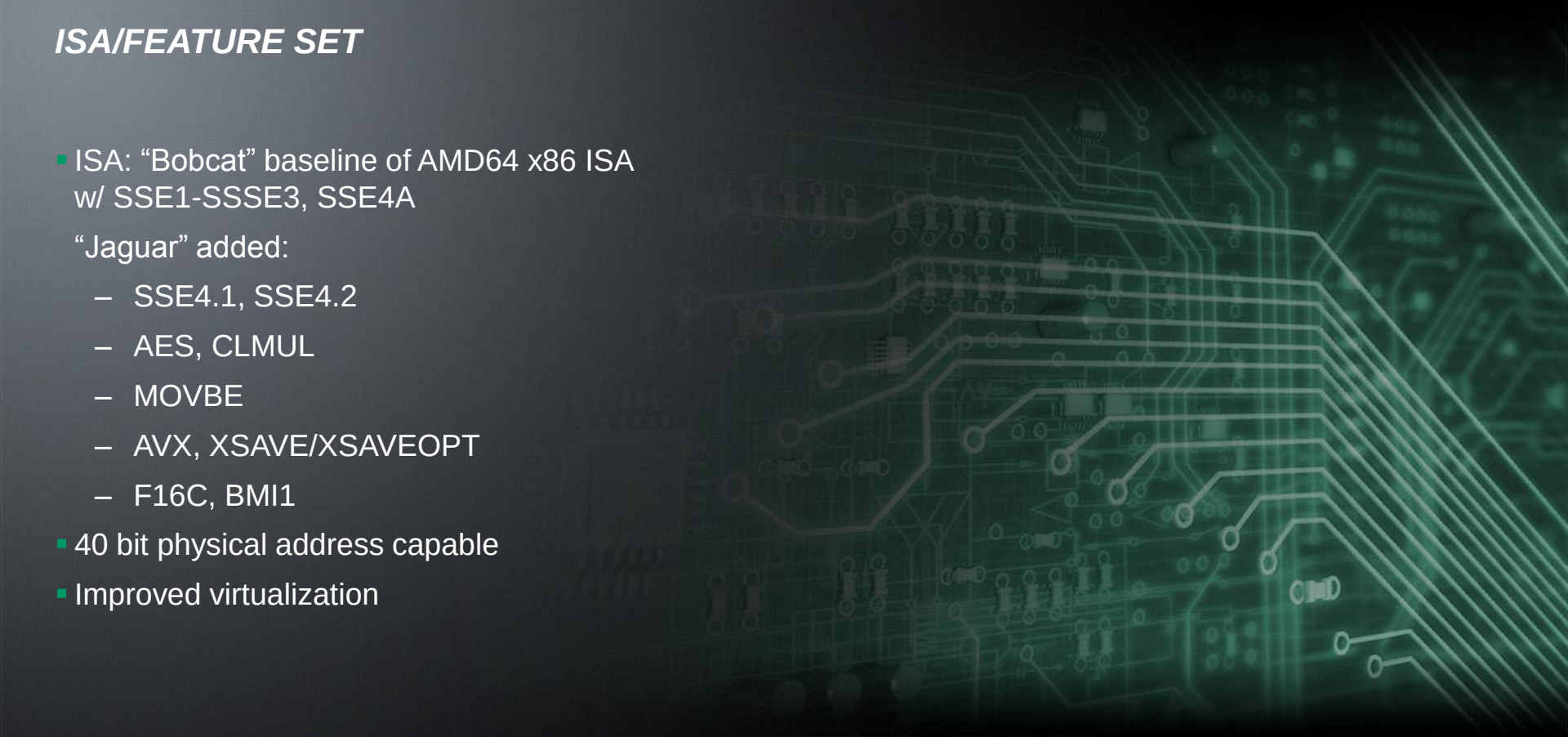


AMD

ISA/FEATURE SET

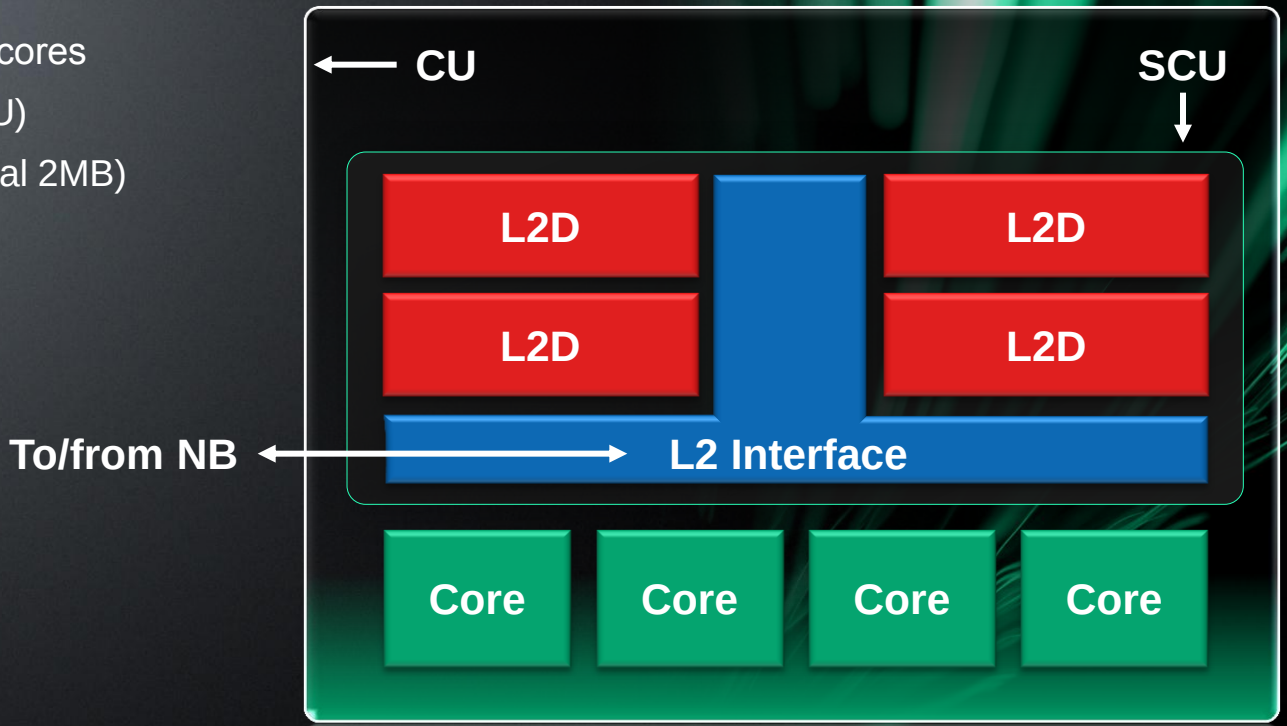
- ISA: “Bobcat” baseline of AMD64 x86 ISA w/ SSE1-SSSE3, SSE4A

 “Jaguar” added:
 - SSE4.1, SSE4.2
 - AES, CLMUL
 - MOVBE
 - AVX, XSAVE/XSAVEOPT
 - F16C, BMI1
- 40 bit physical address capable
- Improved virtualization

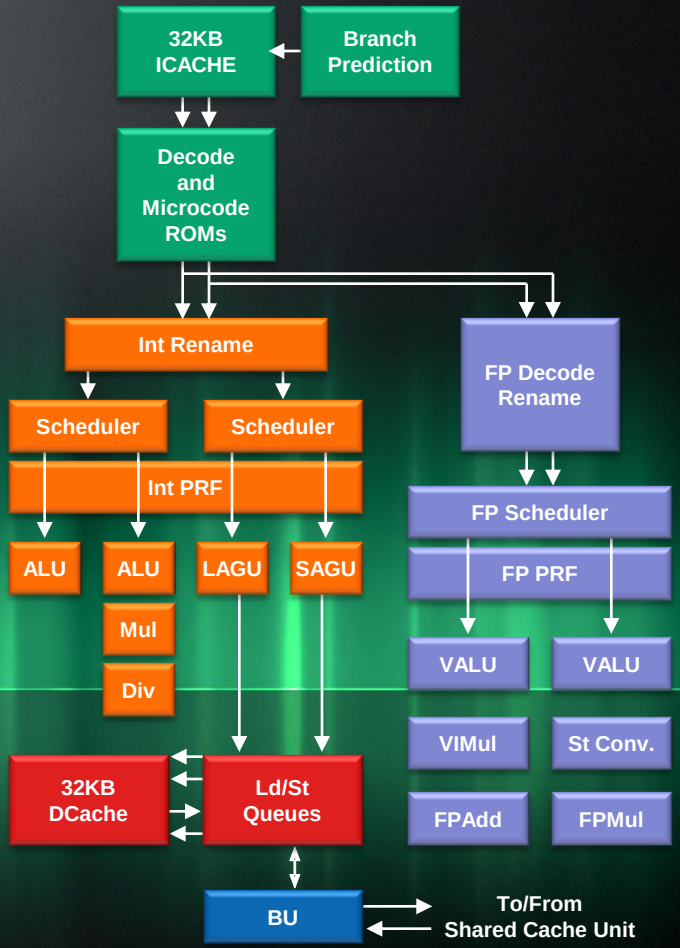


"JAGUAR" COMPUTE UNIT (CU)

- 4 Independent “Jaguar” cores
- Shared Cache Unit (SCU)
 - 4 L2 Data Banks (total 2MB)
 - L2 Interface Tile



"JAGUAR" CORE
Micro-Architecture

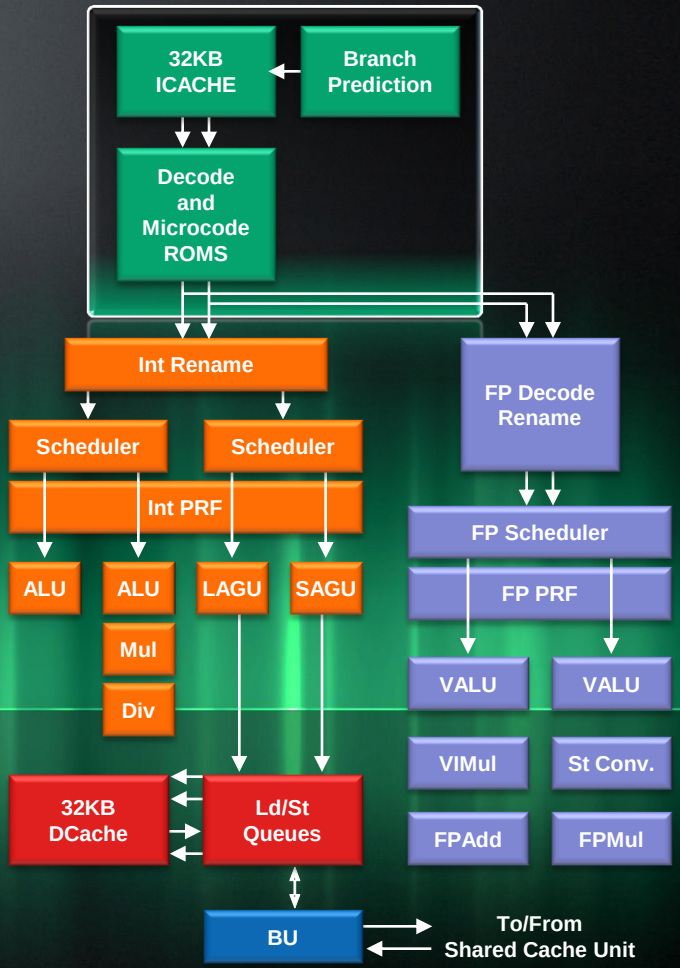


AMD

"JAGUAR" CORE
Frontend

Like “Bobcat”:

- IC: 32KB, 2way
- Itlb: 512 4KB pages
- Layered branch predictor w/ state of the art conditional predictor
- 32B fetch
- 2-instruction decode



“Jaguar” Enhancements:

- 4x32B IC loop buffer for power
- Improved IC prefetcher for IPC
- Grew IB for improved fetch/decode decoupling
- Added decode stage for frequency

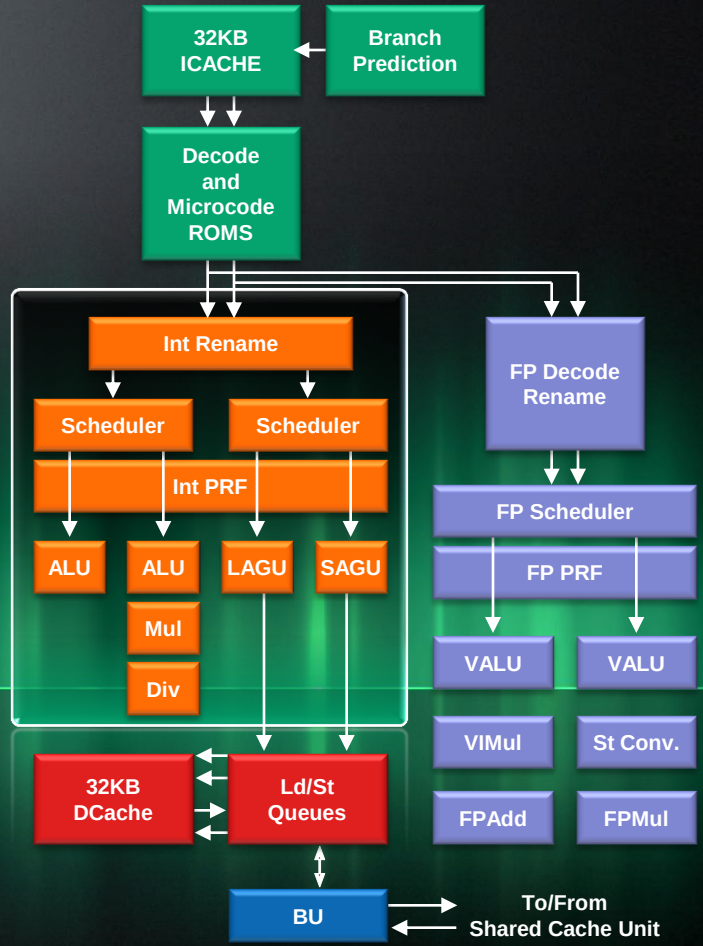
AMD

"JAGUAR" CORE
Integer Execution

Like “Bobcat”:

Schedulers can issue

- 2 ALU
- 1 LD AGU
- 1 ST AGU



“Jaguar” Enhancements:

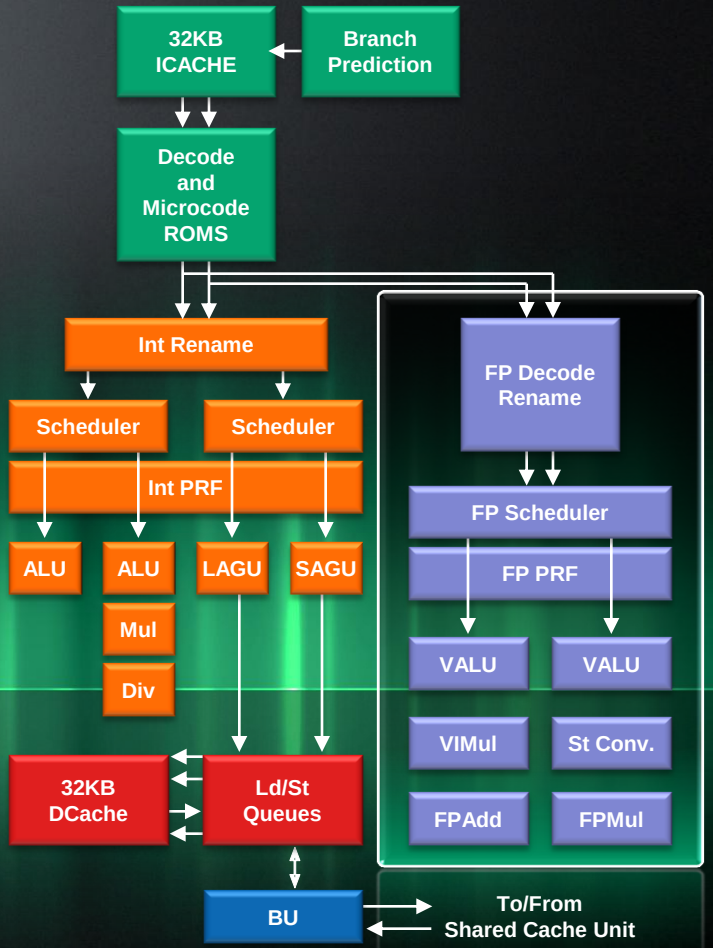
- New hardware divider (leveraged from Llano)
- New/improved cops: CRC32/SSE4.2, BMI1, POPCNT, LZCNT
- More OOO resources Larger schedulers, ROB

AMD

"JAGUAR" CORE
Floating Point Unit

Like “Bobcat”:

- 2 wide FP decode
- OOO scheduler
- 2 execution pipes



“Jaguar” Enhancements:

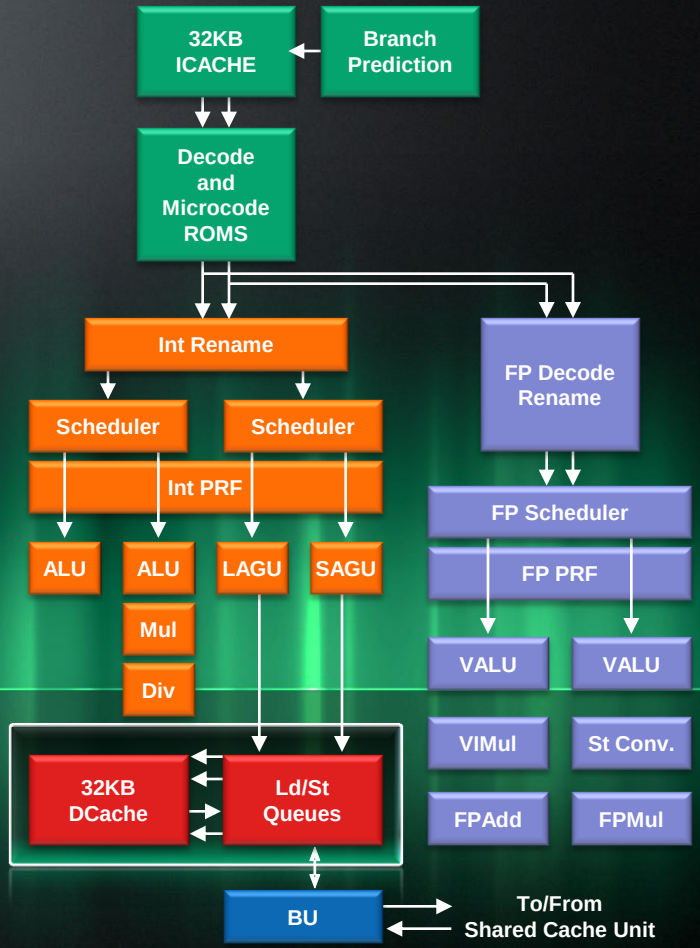
- 128b native hardware
 - 4 SP muls + 4 SP adds
 - 1 DP mul + 2 DP adds
- ISA: many new COPs
 - 256b AVX supported by double pumping 128b hardware
- New Zero Optimizations
- Second FPRF stage for frequency

AMD

"JAGUAR" CORE
Load/Store, Data Cache

Like “Bobcat”:

- DC: 32KB, 8way
- L2DTLB: 512 4KB pages
- 8-stream DC prefetcher
- OOO LS



“Jaguar” Enhancements:

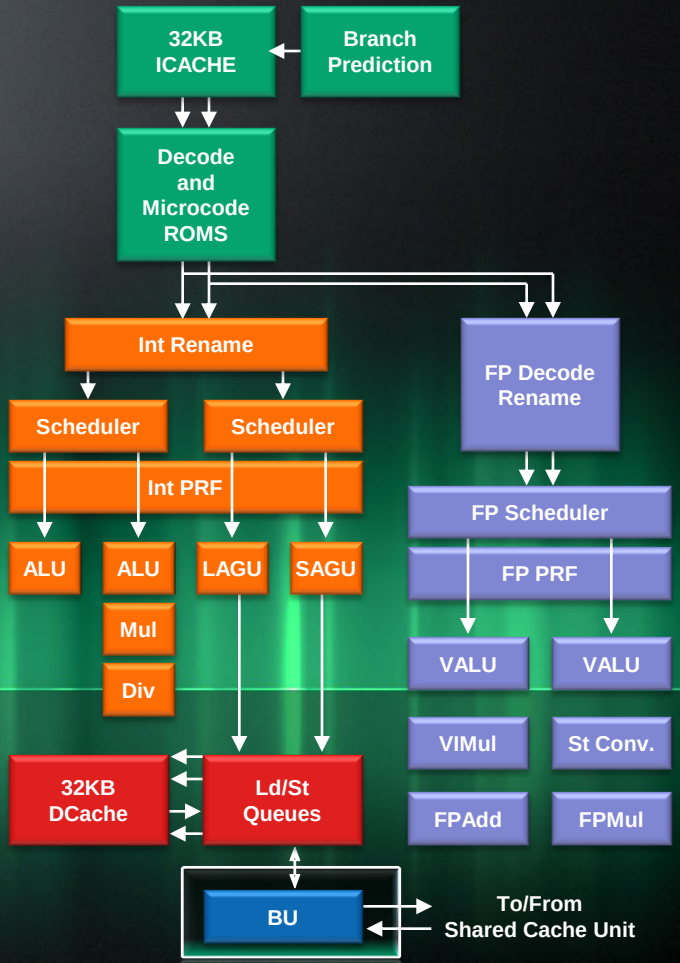
- Ld/St Queues redesign:
 - Improved OOO picker
 - Improved STLF
 - Less store data shuffling
 - More OOO resources
- Enhanced Tablewalks
- 128b data path to FPU

AMD

"JAGUAR" CORE
Bus Unit

Like “Bobcat”:

- BU interfaces between Core (I$,D$) and L2$/NB

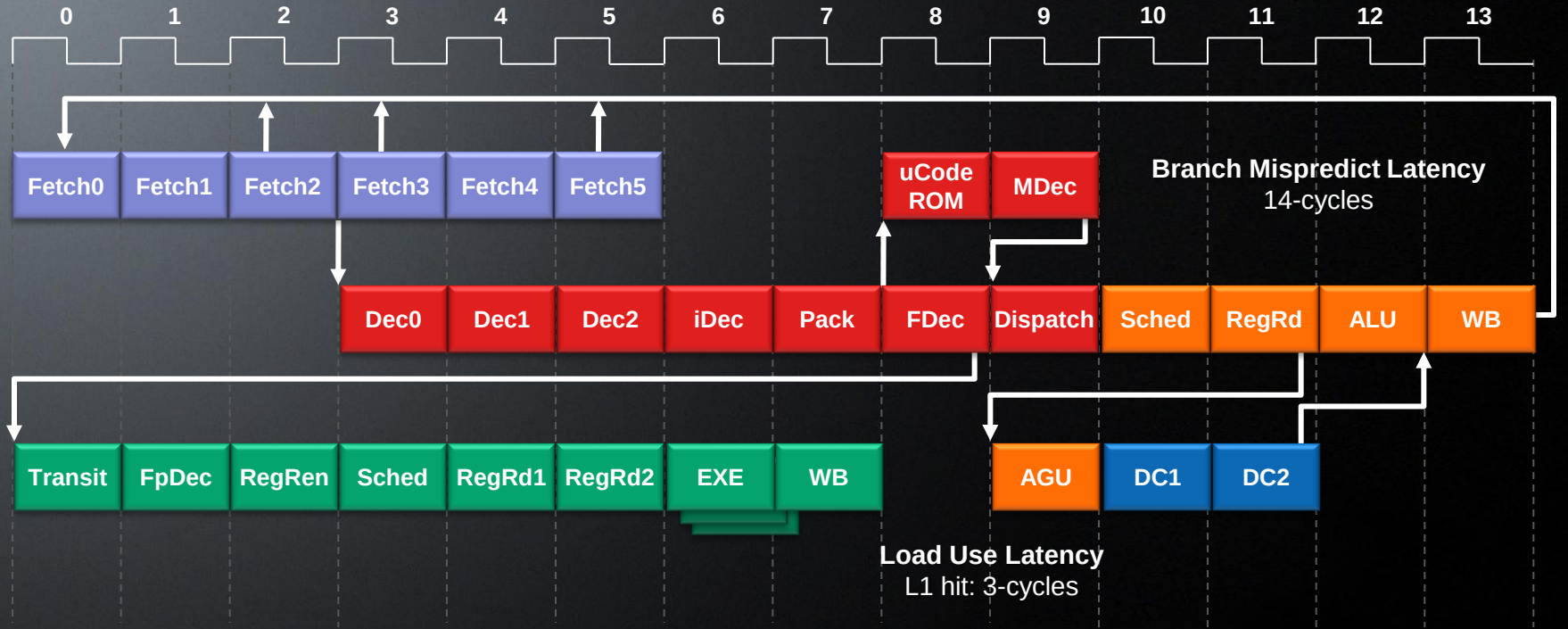


“Jaguar” Supports:

- 8 DC miss/prefetch
- 3 IC miss/prefetch
- Improved Write Combining with 4 WCB data buffers

AMD

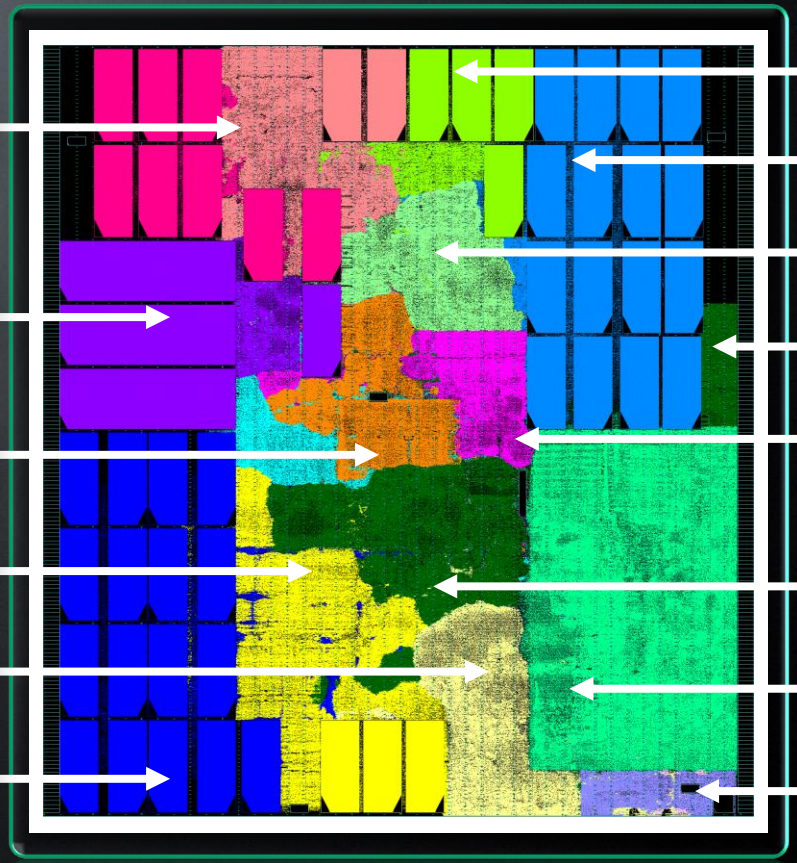
"JAGUAR" CORE PIPELINE



Microarchitectural Frequency improvement over “Bobcat”: >10%

One additional cycle branch mispredict latency vs. “Bobcat”

"JAGUAR" CORE FLOOR PLAN



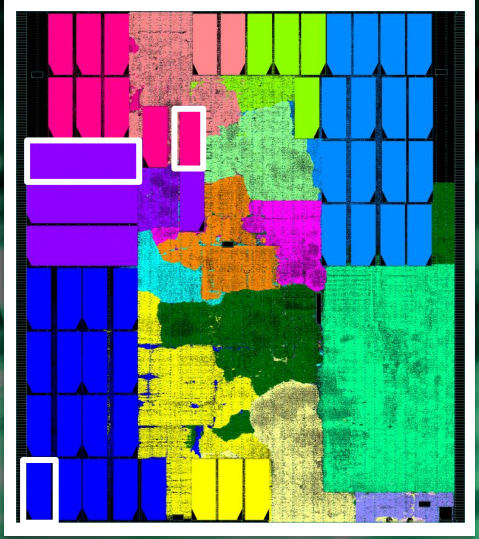
- Branch Predict
- Ucode ROM
- Integer
- Data Tag/TLB
- Bus Unit
- Data Cache
- Inst. Tag/TLB
- Instruction Cache
- x86 Decode
- Debug
- ROB
- Load/Store
- FP
- Test

AMD

CORE FLOOR PLAN COMPARISON



“Bobcat” core in 40nm = 4.9 mm^2
7 core macros, 2 L2 macros,
3 clock macros



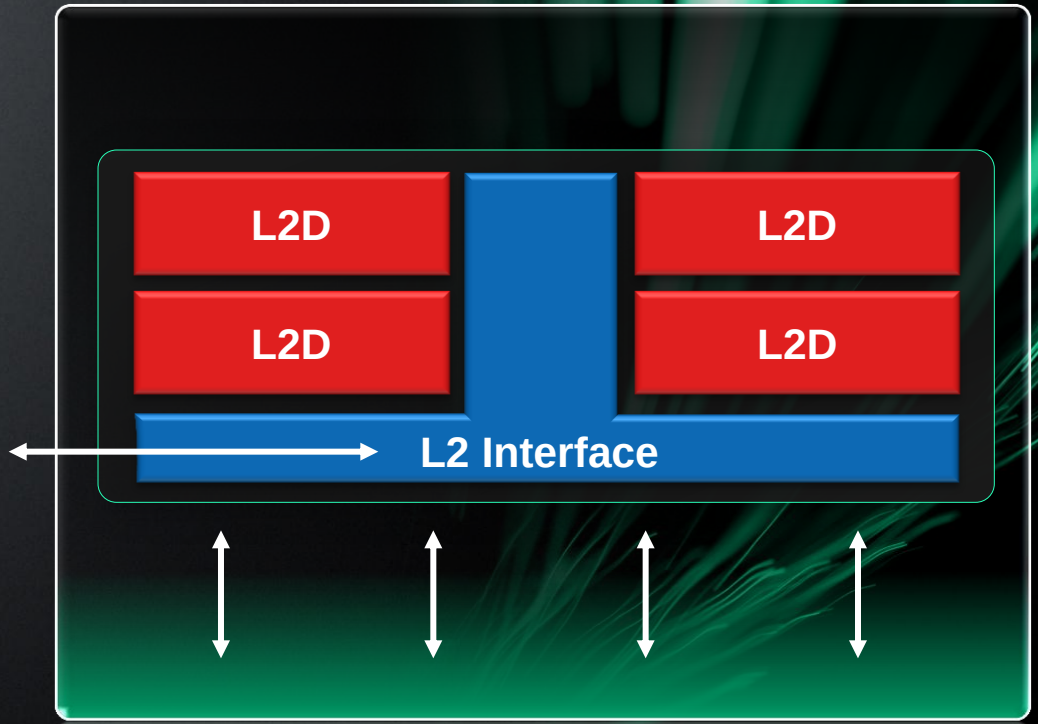
“Jaguar” core in 28nm = 3.1 mm^2
3 core macros, 1 L2 macro,
1 clock macro

AMD

"JAGUAR" SHARED CACHE UNIT

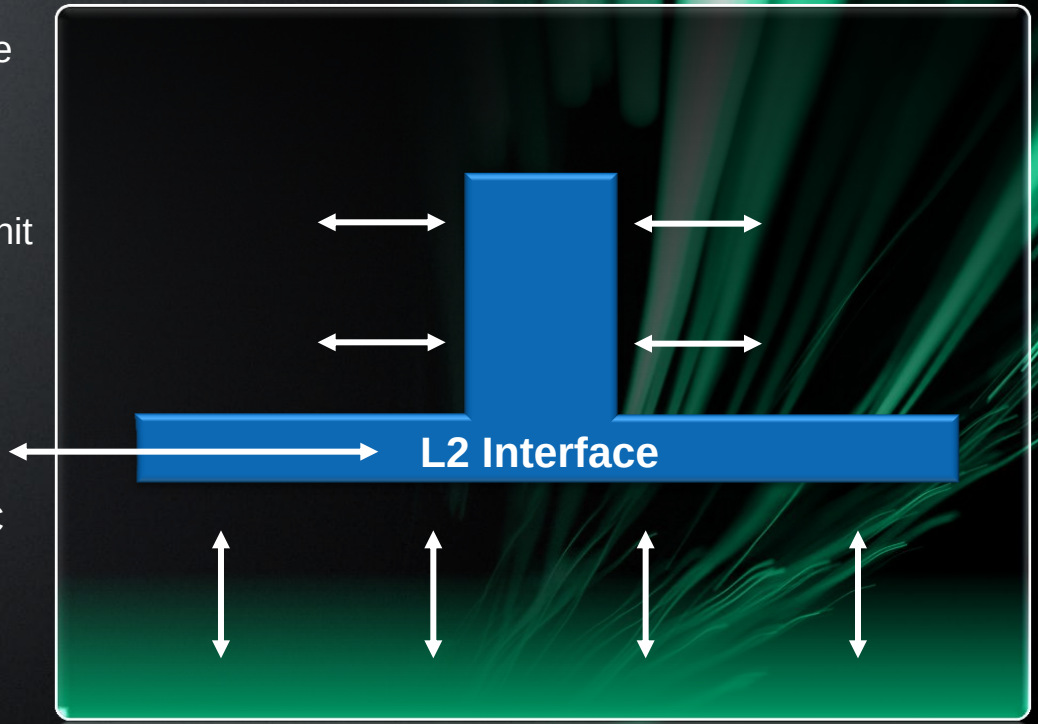
Shared Cache is major design addition in “Jaguar”

- Supports 4 cores
- Total shared 2MB, 16-way
 - Supported by 4 L2D banks of 512KB each
- L2 cache is inclusive – allows using L2 tags as probe filter
 - Any line in a Core L1 instruction or data cache must be in the L2



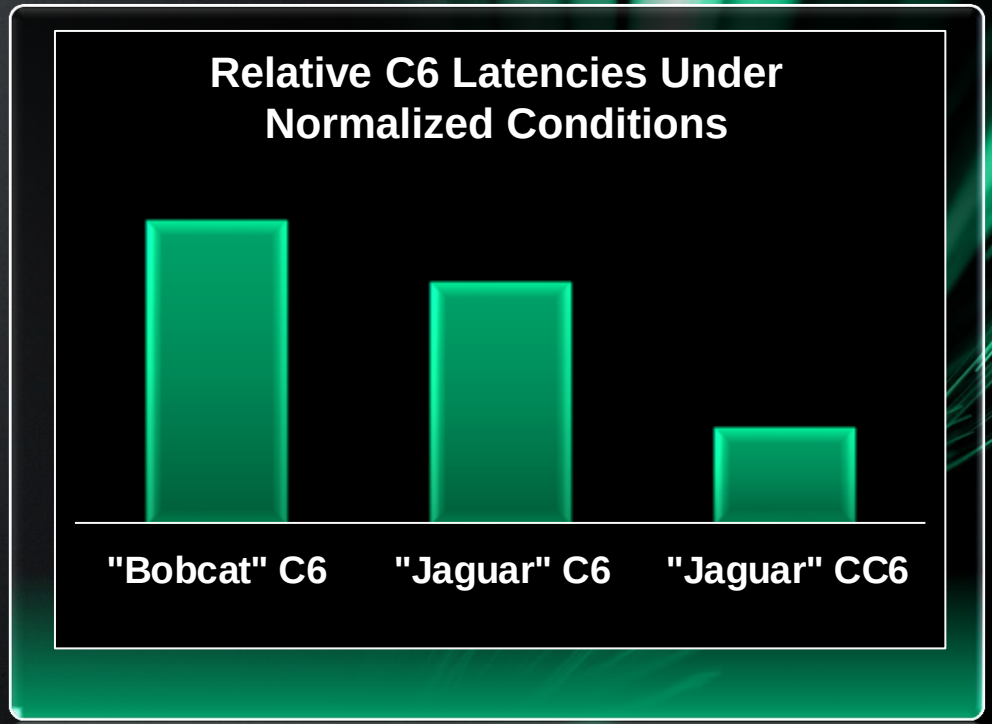
"JAGUAR" L2 INTERFACE

- All connections routed thru L2 interface
- L2 tags reside in interface block
 - Divided into 4 banks
 - L2D bank lookup only after L2 tag hit
- L2 Interface block runs at core clock
 - L2D's run at half clock for power, only clocked when required
- New L2 stream prefetcher per core
 - Allows improved bandwidths & IPC
- Up to a total of 24 paired read + write transactions in flight
- 16 additional L2 snoop queue entries
 - Allows for handling coherent probes at high bandwidth



"JAGUAR" C6

- Any Core can independently go into CC6 power gating
 - Optimized microcode routines and hardware allow for fast CC6 entry/exit
 - Shared L2 leaves more cache for the remaining active cores (IPC)
- Last core in the compute unit to be power gated flushes shared L2 in preparation for full C6. Hardware engines added to improve L2 flush times.



"JAGUAR" POWER

- Many blocks redesigned for improved power efficiency
 - IC Loop Buffer, Store Queue, L2 clocks, etc.
- Clock power usage scrubbed, including improved dynamic clock gating:

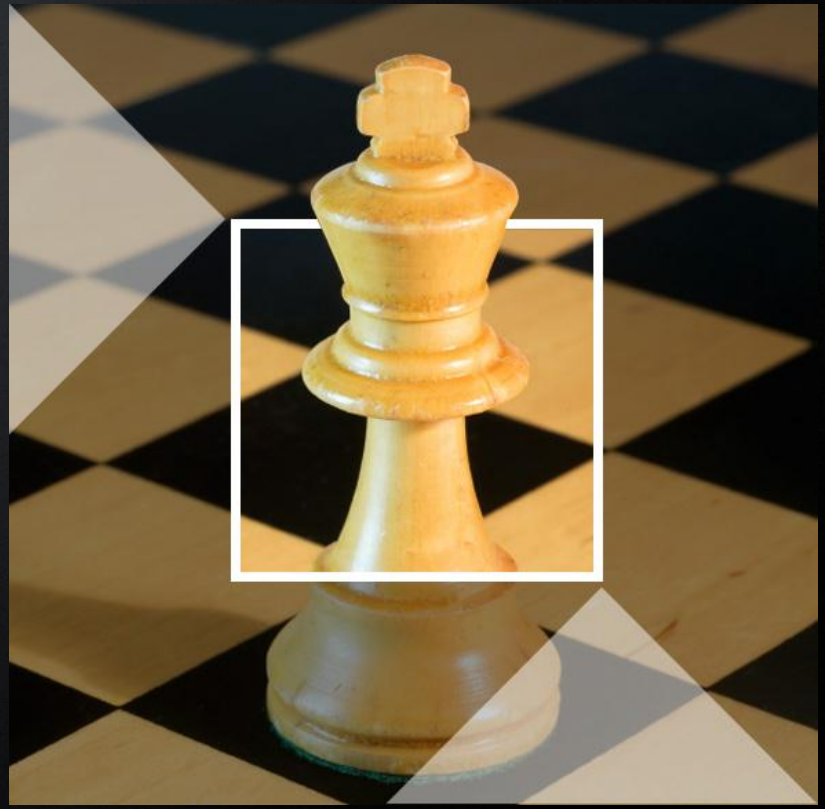
	"Bobcat" IPC	"Jaguar" IPC	"Bobcat" Core Gater Efficiency	"Jaguar" Core Gater Efficiency
Halt	0.00	0.00	91.8	98.8
Apps	0.95	1.10	89.7	92.3
"Bobcat" Virus	1.74	1.78	84.6	87.1
"Jaguar" Virus	0.81	1.86	85.7	85.0

- Increased frequency capability allows choices:
 - Higher frequency -> higher performance
 - Same frequency at lower voltage -> lower power

AMD

SUMMARY

- ISA enhancements
- Increased process portability
- Estimated typical IPC improvement: >15%*
- Frequency improvement: >10%*
- Dynamic power efficiency improvements



*Based on internal AMD modeling using benchmark simulations

AMD

DISCLAIMER & ATTRIBUTION

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. There is no obligation to update or otherwise correct or revise this information. However, we reserve the right to revise this information and to make changes from time to time to the content hereof without obligation to notify any person of such revisions or changes.

NO REPRESENTATIONS OR WARRANTIES ARE MADE WITH RESPECT TO THE CONTENTS HEREOF AND NO RESPONSIBILITY IS ASSUMED FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

ALL IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE ARE EXPRESSLY DISCLAIMED. IN NO EVENT WILL ANY LIABILITY TO ANY PERSON BE INCURRED FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

Trademark Attribution

©2012 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, Phenom, Radeon, and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. Microsoft and DirectX are registered trademarks, of Microsoft Corporation in the United States and/or other jurisdictions. Other names used in this presentation are for identification purposes only and may be trademarks of their respective owners.



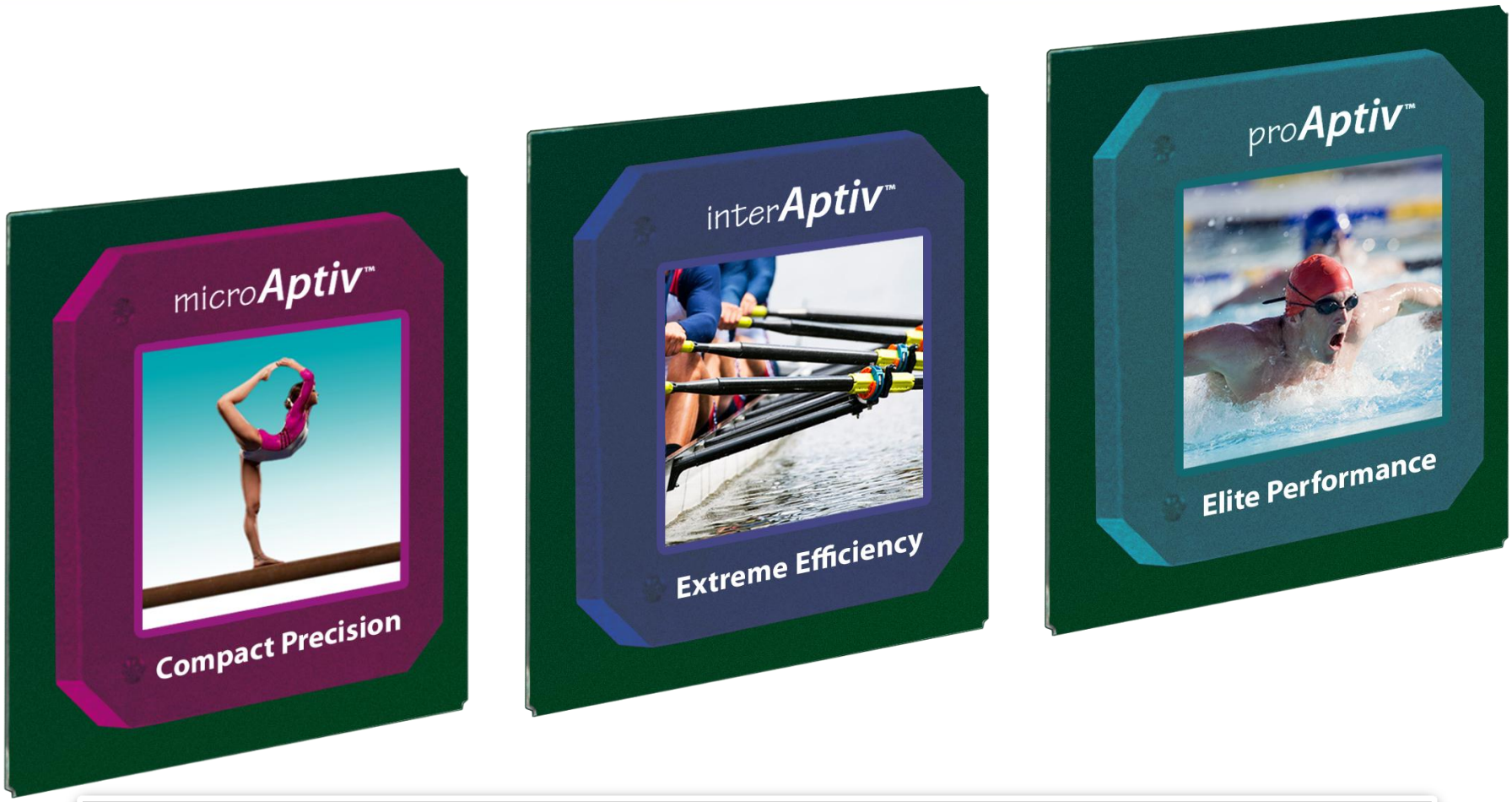
proAptiv: ***Efficient Performance*** on a Fully-Synthesizable Core

28 August 2012

Ranganathan "Suds" Sudhakar
Chief Architect

© 2012 MIPS Technologies, Inc. All rights reserved.

Aptiv Family Highlights



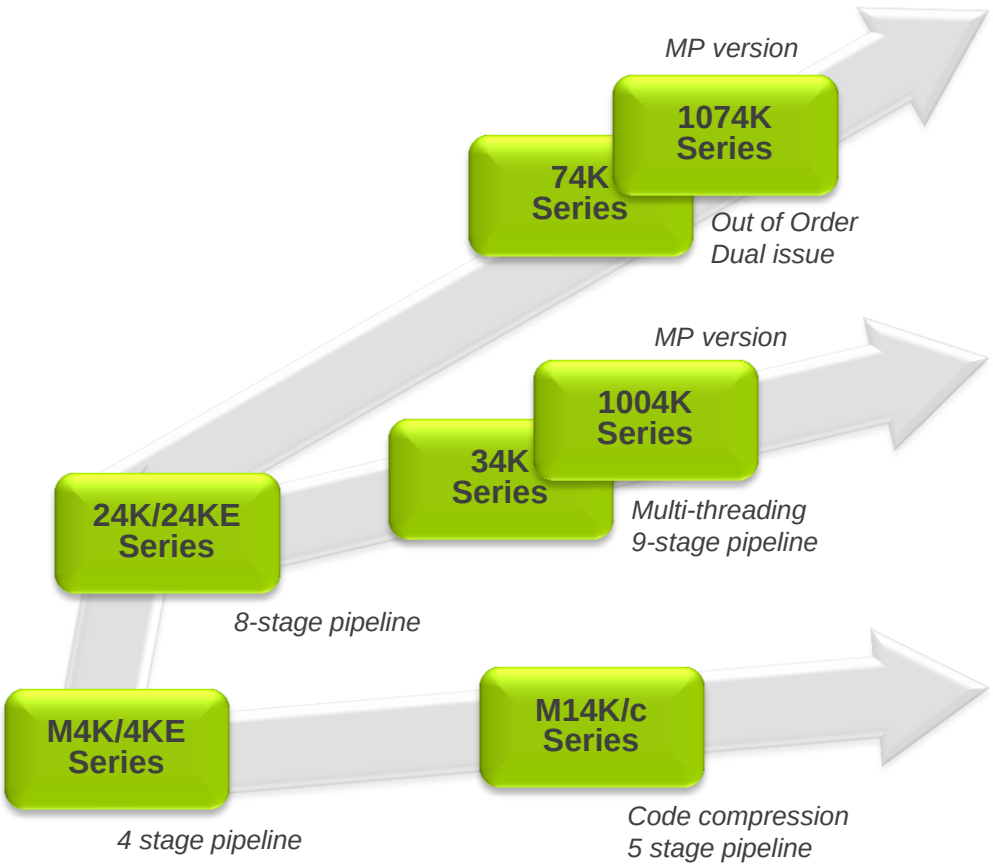
Three new cores optimized for embedded markets

© 2012 MIPS Technologies, Inc. All rights reserved.

Aptiv Core Portfolio

Classic MIPS Products

Aptiv™ Generation ⟶



proAptiv™ Family

1 to 6 core configs,
Hi-speed FPU and
L2 cache controller

Single-Threaded Area Optimized

interAptiv™ Family

1 to 4 core configs,
2-level MT/FPU and
L2 cache controller

Multi-Threaded Power Optimized

microAptiv™ Family

MCU (cacheless) or
MPU (caches/TLBs)
with real-time/security

DSP-Accelerated Energy Optimized

© 2012 MIPS Technologies, Inc. All rights reserved.

What is a “Soft” Core?

- **Fully synthesizable “package”**
 - Design data
 - RTL
 - Configurator – MP/MT, FPU, Trace/Debug, cache/TLB/SPRAM/buffer sizes, bus widths
 - Physical design support
 - Reference floorplans, Synthesis + Place-and-Route scripts
 - DFT/Scan, Timing and Power Analysis scripts
 - Simulation models
 - Bus Functional Models and compliance checkers
 - Instruction accurate simulators, Cycle exact simulators
 - Verification collateral
 - Architectural Verification Test suites, core diagnostics
 - Sample testbench, build and run scripts
 - Documentation
 - ISA manuals, global configuration register tables, memory maps, boot procedures
 - Implementer’s Guide, Integrator’s Guide, Hardware/Software User manuals

- **Available separately**
 - FPGA development boards
 - EJTAG/debug probes
 - OS components, libraries, software toolchains (compiler, libraries, JITs, codecs)

© 2012 MIPS Technologies, Inc. All rights reserved.

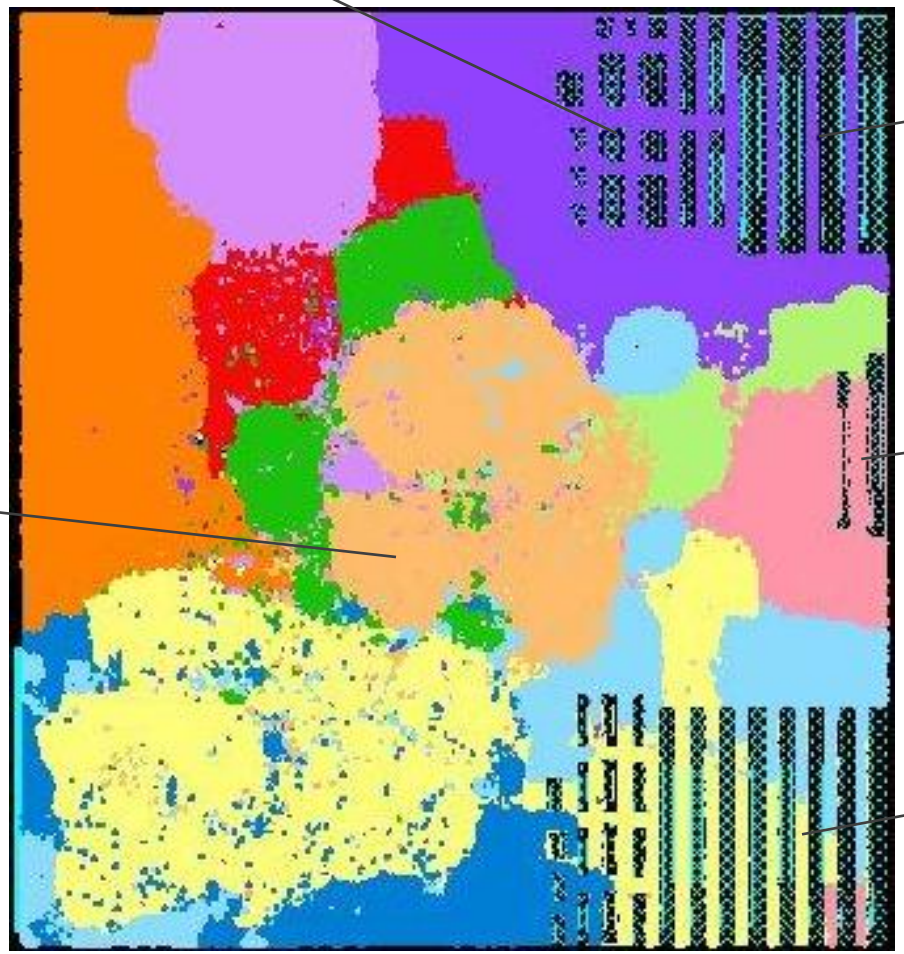
What is a “Hardened” Core?

- **Tapeout-ready GDS, built on a generic ASIC flow, using:**
 - Configured soft core
 - Floorplan – placement of RAMs, bounding box
 - Physical IP for some process technology
 - Standard cell library
 - e.g. 28nm low-leakage 12-track mixed-Vt with booster flops
 - Compiled memories
 - e.g. 28nm high-speed LVT single + dual-port bit-writable memories
 - Fab conditions
 - Process corner **(usually worst-case slow-slow, high-temp, low voltage)**
 - Number of metal layers, DRC/LVS, power grid, IR drop, OCV/AOCV, PLL jitter

- **Not to be confused with a “hard core”**
 - Frequency and power improvements beyond simple hardening:
 - Custom std cells, flops, clk-gaters characterized for typical silicon
 - e.g. 1.x GHz worst-case SVT → 2.x GHz typical with LVT, overdrive, cooling
 - Multi-port register files and custom memories for cache arrays
 - Hierarchical floorplans, structured placement, mesh clocking

© 2012 MIPS Technologies, Inc. All rights reserved.

Hardened proAptiv Layout



© 2012 MIPS Technologies, Inc. All rights reserved.

Soft Core Design Considerations

- **Life revolves around flops (and muxes)**
 - No CAMs – schedulers, TLBs, BTBs all built from flops
 - No ROMs – div/sqrt lookup tables all built from gates
 - No multiports – Register files, reorder buffers all built from flops
 - Read ports are large muxes ~O(num_entries)
 - Write ports are small muxes ~O(num_ports)
 - Exceptions are:
 - 1RW RAMs for use in cache/TLB arrays
 - 1R1W RAMs for use in branch prediction arrays
 - Used judiciously -- proAptiv is the first MIPS soft core to use these
- **Sophisticated techniques cannot be easily employed**
 - Banking, sum-addressing or one-hot-indexing
 - Dynamic circuits, especially negedge-triggered
- **More pipestages needed for a given frequency**
 - MIPS's pure RISC ISA helps counteract this

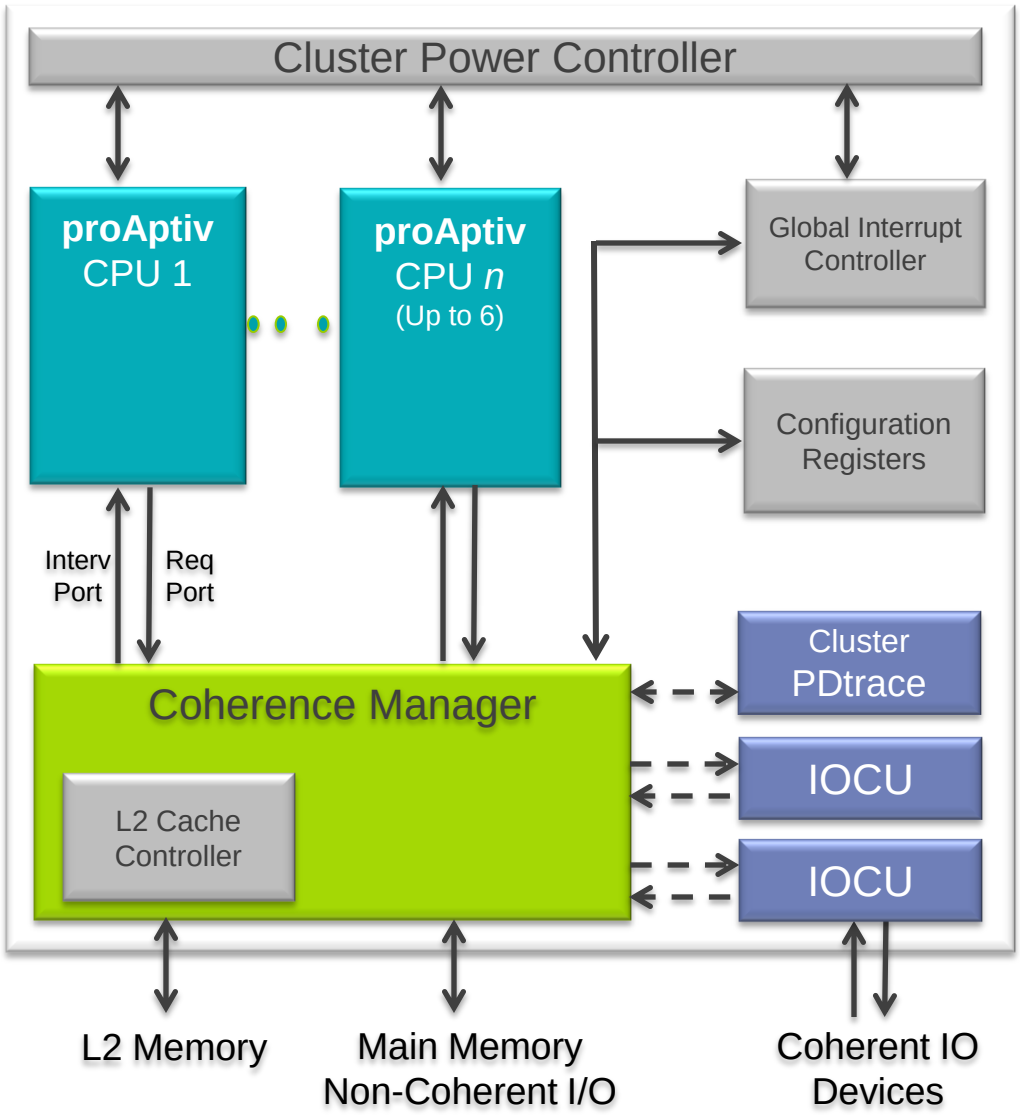
© 2012 MIPS Technologies, Inc. All rights reserved.

Soft Core Timing and Verification Challenges

- **Timing paths not consistent**
 - Variations in floorplan, configuration, stdcell, memory IP
 - Variations in operating point – fab, process, Vt mix, overdrive
 - Variations in EDA tool margins, flows, vendors and versions
- **But good enough!**
 - Balance logic across pipestages
 - Ensure loop paths are minimal and reflected in the microarchitecture
 - Ensure floorplan reflects critical unit and pin placement
- **Specific considerations for high-frequency pipelines**
 - Any CAM-RAM structures take at least 2 clock cycles
 - Regfile read+bypass takes at least 2 clock cycles
- **Need to fix timing paths at all phases of the implementation**
 - Synthesis, Place, Route, Clocking (No ECOs or manual tuning allowed)
- **Verification challenges**
 - Dozens of configuration variables but still need high code+functional coverage

© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Coherent Processing System (CPS)



- **Modern Superscalar Core**
- **Enhanced Coherence Manager**
 - Integrated L2$ Controller with ECC
 - Improved CM performance
 - Supports configurations up to 6 cores
- **IO Coherence Unit (IOCU)**
 - Up to 2 IOCU Blocks
- **Global Interrupt Controller**
 - Up to 256 system interrupts
- **Cluster Power Controller**
 - Voltage domain/gating per core
 - Clock gating per core
 - Software programmable
- **PDtrace™ – cluster level support**

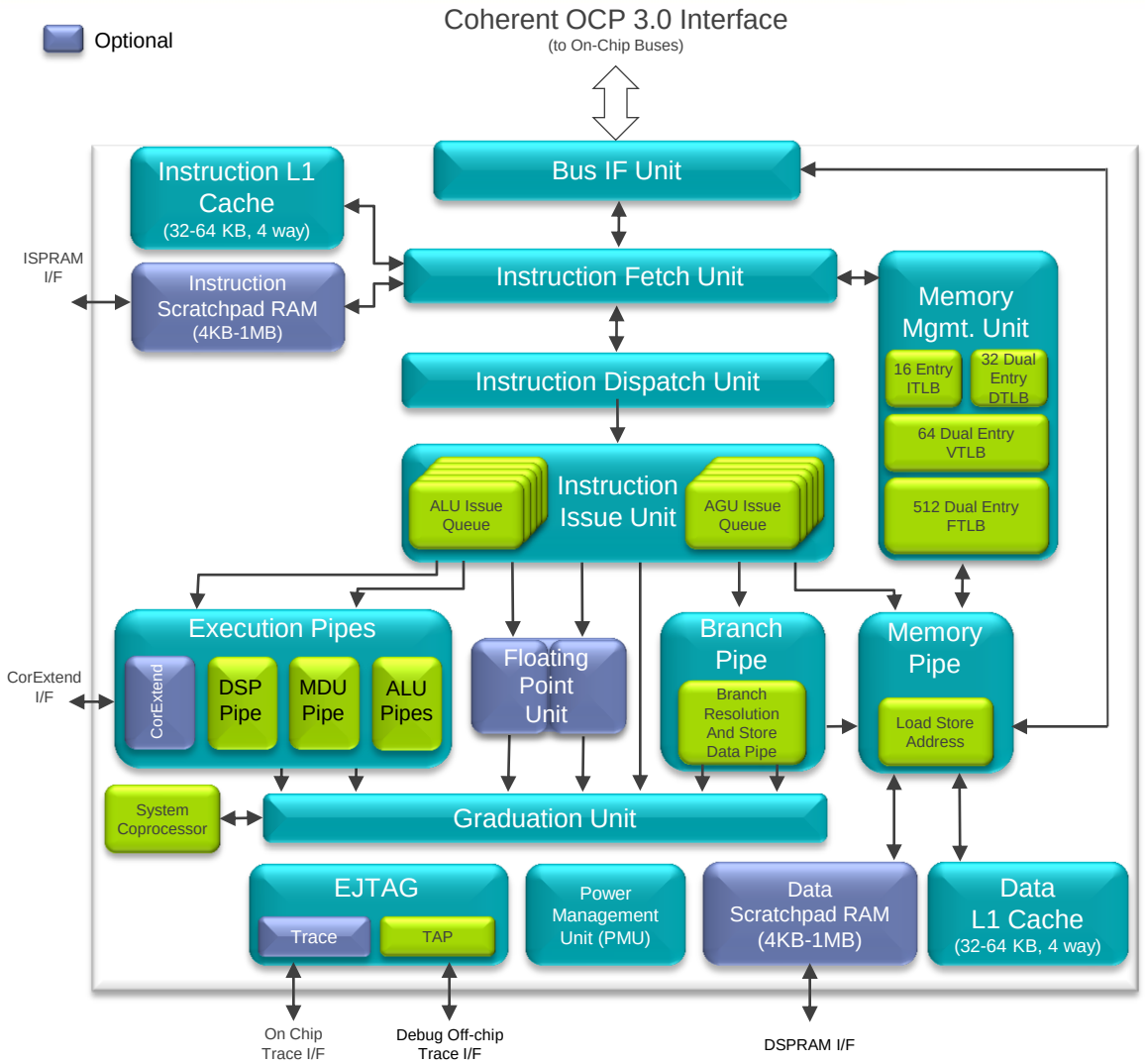
© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Design Goals

- **Fast**
 - Optimized for mobile computing and networking
 - Multi-issue dynamically-scheduled operation
 - Deep pipeline to achieve multi-gigahertz operation
 - Brand new high-frequency FPU matched to core
- **Efficient**
 - Elegant balanced microarchitecture, not brute force width and depth
 - Minimal area for cost and leakage; fine-grain clock gating
 - Reduces the need for costly heterogeneous schemes
- **Scalable**
 - New 6-core Coherence Manager and 256-bit L2 cache controller
- **Robust**
 - Age-based scheduling, careful tuning of predictors/prefetchers
 - Easy to add features and performance, vary microarch parameters
- **Feature set**
 - MIPS32 R3 / MIPS16e, DSP ASE v2, PDtrace v6, Enhanced VA

© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Base Core Architecture

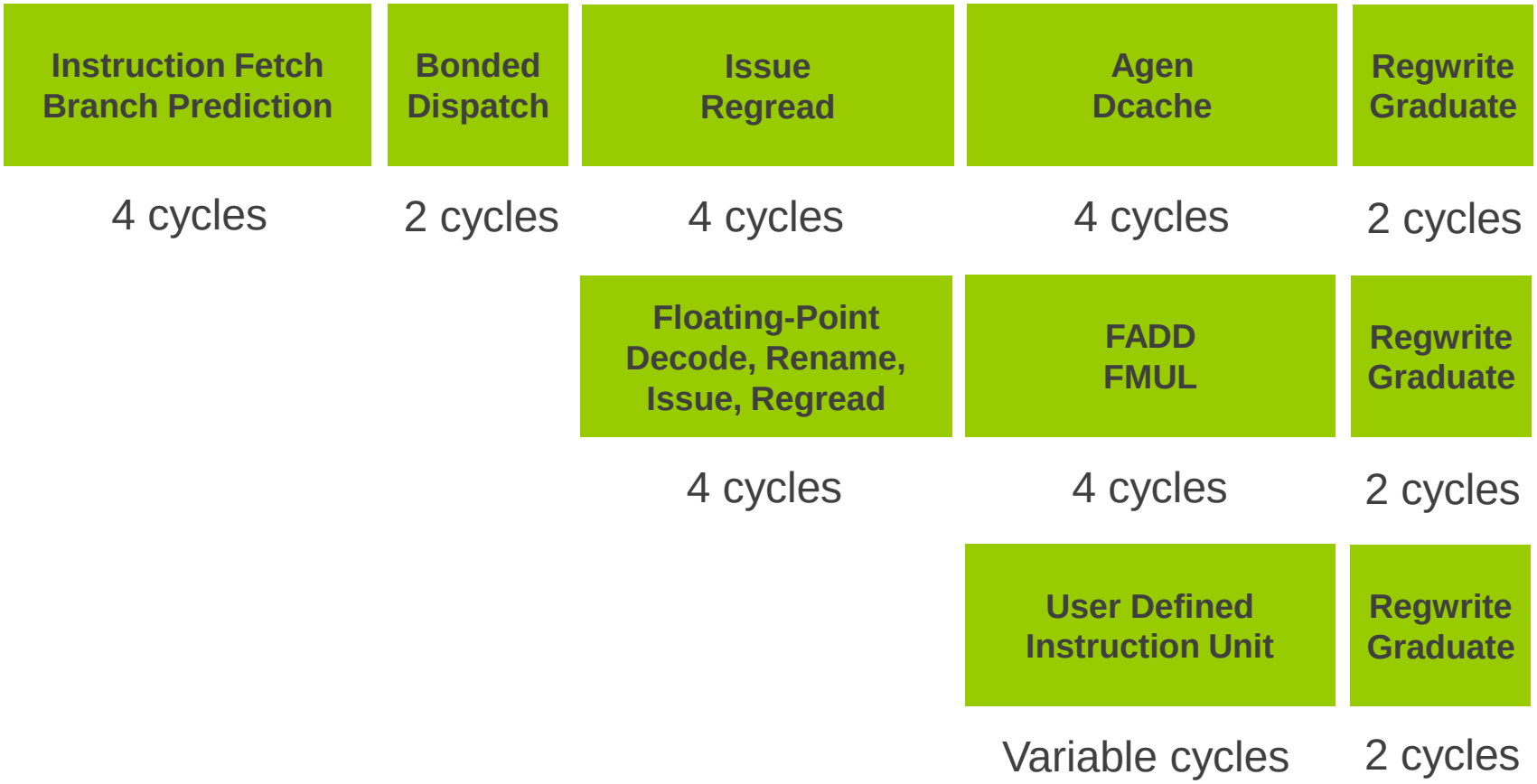


- **Superscalar OoO CPU – 16 stage**
 - Quad inst fetch
 - Triple bonded dispatch
 - Inst peak issue: quad integer; dual FP
- **Sophisticated branch prediction and L0/L1/L2 BTBs, RPS, JRC, way predicted instruction cache**
- **High performance, multi-level TLBs, way predicted data cache**
- **Instruction Bonding makes six issue pipes look like eight**
- **Fast integer divide, multiply and multiply-accumulate operations**
- **Dual Issue FPU**
 - Higher speed (1:1 with CPU)
 - Lower latency on most operations
 - Single-pass double precision
 - More parallelism and dedicated schedulers – more ops in flight

© 2012 MIPS Technologies, Inc. All rights reserved.

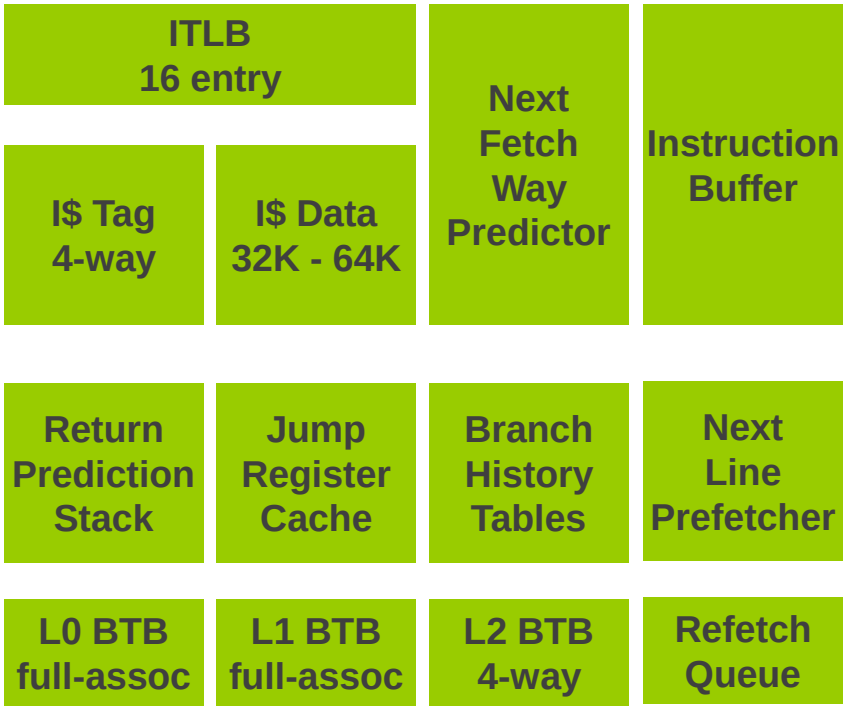
proAptiv Pipeline

16 stage integer load pipeline



© 2012 MIPS Technologies, Inc. All rights reserved.

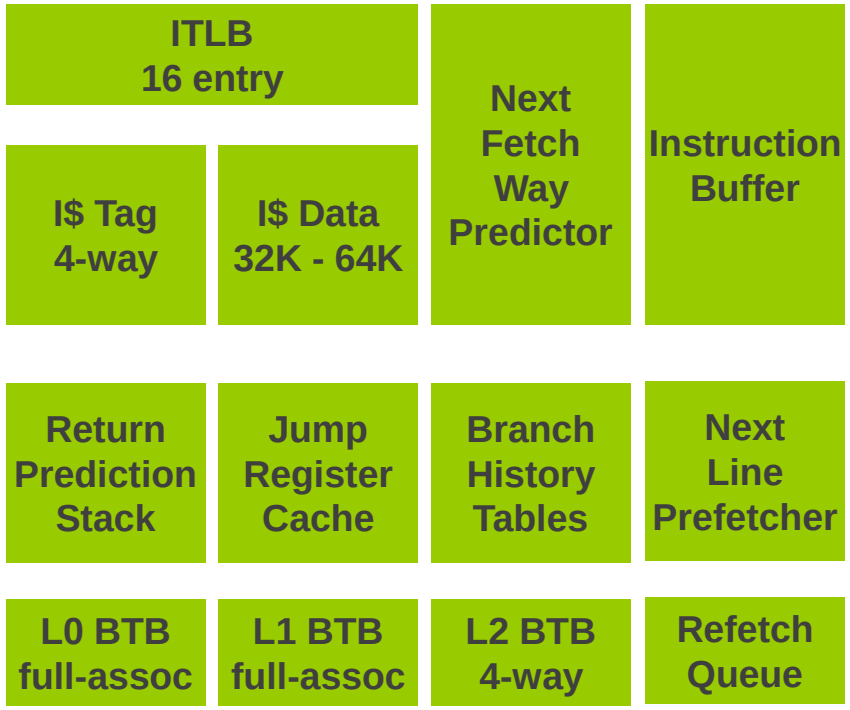
proAptiv Instruction Fetch



- 16-entry ITLB
- 32 or 64KB I-cache
- 4-way associative LRU
- 32-byte line
- Parity protected
- Fetches 16 bytes per cycle
 - Aligned fetch reduces power
 - Dynamic 8 byte bundle reduces power on branches or MIPS16e
- Next Fetch Way predictor reduces power by reading only one way
 - Sequential fetch way from SRAM
 - Target way from BTB
- Next Line Prefetcher
 - Variable number of lines on a miss
 - Direct bypass from refetch queue
- Extra pipestages inserted for MIPS16e fetching and unpacking
- Credit-based instruction buffer

© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Branch Prediction



- **Branch History Tables**
 - Predicts branch direction
 - Novel algorithms deliver class-leading prediction accuracy
 - Uses sophisticated global history
 - Can predict 2 (MIPS32) or 4 (MIPS16e) branches per cycle
 - Multiple SRAM-based tables
 - Only 1R1W structures on chip
 - Leverages delay slots to minimize storage capacity needed
 - In MIPS, unlike some ISAs, the delay slot cannot itself be a branch
- **Branch Target Buffers**
 - Provides fast target prediction
 - Multiple buffers with various latencies and sizes, up to 512 entry 4-way
- **Jump Register Cache**
 - Predicts indirect jumps
 - Multiple targets per jump PC
- **Return Prediction Stack**

© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Instruction Dispatch – Bonding

- **Combine adjacent instructions into single bonded op**
 - e.g. consecutive LW or SW instructions
 - e.g. branch with certain instructions in delay slot
 - Fused compare-branch is already part of MIPS integer ISA
- **Load/Store bonding makes one memory pipe look like two**
 - 1 DTLB, 1 tag array, single-ported data array saves area
 - Single DTLB and cache access saves energy, power
 - Occupies only 1 entry in various queues/buffers – more ILP
 - Carried forward as one operation on L1-miss – more MLP
 - Speeds memset, bcopy, strcmp, spill-fill, GPU communication
- **Design decisions**
 - Initially limit to two instructions, aligned addresses and ST
 - But designed to scale to four, misaligned accesses and MT
 - Therefore, needs a *bonding predictor* in the front-end
 - Trained by LSU (memtype must be cacheable or write-combining)
 - Indexed by PC and other control flow information

MemCopy Loop:

```
lw    r1, 0x0(r20)
lw    r2, 0x4(r20)
lw    r3, 0x8(r20)
lw    r4, 0xc(r20)
lw    r5, 0x10(r20)
lw    r6, 0x14(r20)
lw    r7, 0x18(r20)
lw    r8, 0x1c(r20)

sw    r1, 0x0(r21)
sw    r2, 0x4(r21)
sw    r3, 0x8(r21)
sw    r4, 0xc(r21)
sw    r5, 0x10(r21)
sw    r6, 0x14(r21)
sw    r7, 0x18(r21)
sw    r8, 0x1c(r21)

addiu  r20, r20, 0x20
addiu  r21, r21, 0x20
bnez   r23, Loop
sub    r23, r23, r22
```


© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Instruction Dispatch – Cracking

- **Bonded stores have 3 source registers**
 - 1 address and 2 data GPRs
 - Compared to 2 sources for ordinary stores
 - Requires 1 more read port at execute than unbonded machine
- **Hence cracked into decoupled operations**
 - STA (Store Address) – 1 reg source
 - STD (Store Data) – 2 reg sources
- **STA reads cache tags and detects L1 miss early**
 - Requires only 1 read port in load-store pipe
- **STD delivers data to LSU in memory aligned format**
 - Requires only 2 read ports
 - Thus avoiding the need for any pipe to have 3 ports
- **Some stores are never cracked**
 - e.g. Misaligned stores, where data *depends* on address
- **Some stores are always cracked**
 - e.g. FPU stores, where the integer scheduler has no visibility or control over the FP register file and issue ports

© 2012 MIPS Technologies, Inc. All rights reserved.

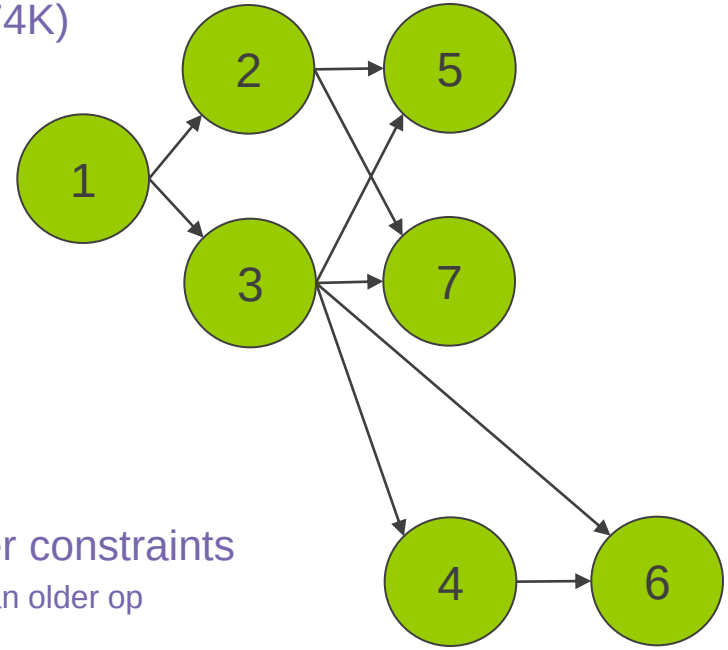
proAptiv Instruction Issue – Segmented Scheduler

- **Two issue queues**
 - Neither single large unified queue (low-frequency)
 - Nor too many small distributed schedulers (high power)
- **1 ALU issue queue and 1 AGU issue queue**
 - Check dependencies and structural hazards
 - STA and STD share same scheduler entry, reducing area/power
- **Age-priority scheduling**
 - Requires age-vector per entry to pick oldest
 - Allows non-shifting schedulers with fewer comparators/muxes for low power
 - Minimal CAM logic – timing friendly
- **No reservation stations**
 - Read registers after scheduling – low power

© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Instruction Issue – Transitive Wakeup

- **Holy grail of OoO scheduler design:**
 - Large (40 – 64 entries) yet fast (able to follow single-cycle dependency chains)
- **Typical schedulers employ one of two wakeup techniques**
 - Encoded register-number wakeup (e.g. MIPS R10K)
 - (Wakeup → Pick → Mux) → (Wakeup → Pick → Mux) → …
 - Pick and Mux can sometimes be overlapped
 - Decoded entry-number wakeup (e.g. MIPS 1074K)
 - (Wakeup → Pick) → (Wakeup → Pick) → …
 - Usually multi-hot vectors for dependency checking
- **proAptiv can utilize a third technique**
 - Transitive Wakeup
 - (Wakeup) → (Wakeup) → (Wakeup) → …
 - Only works with decoded entry-numbers
 - Relies on multi-hot broadcasts
 - {1} → {1, 2, 3} → {1, 2, 3, 4, 5, 7} → {1, 2, 3, 4, 5, 6, 7}
 - Requires strict age-priority scheduling and other constraints
 - Prevents premature pick of a younger op dependent on an older op
 - e.g. inst 6 before inst 4



© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Integer Execution

- **One simple ALU pipe**
 - Handles arithmetic, logical ops and small shifts
- **One complex ALU pipe**
 - Handles a superset of the simple ALU ops – such as large Shifts
 - Handles DSP operations that involve reading or writing the 64b accumulators
 - Accumulators are renamed and treated as two 32b registers
 - Saves power and area compared to designs using 64b rename pool
 - DSP flags are renamed using separate 13b wide pool
 - Allows easy handling of sticky status bit fields
 - Interfaces with Multiply-Divide Unit which also uses the accumulators
 - Supports single-cycle bypass for integer multiply-accumulate
 - New designs for fast multiplication and very fast division
- **One branch/store-data pipe**
- **One load/store pipe**
- **Pipes share read and write ports to further bring down area/power**

- **Thanks to bonding, the 4 physical pipes can actually execute up to 6 MIPS32 integer instructions on a particular clock cycle**

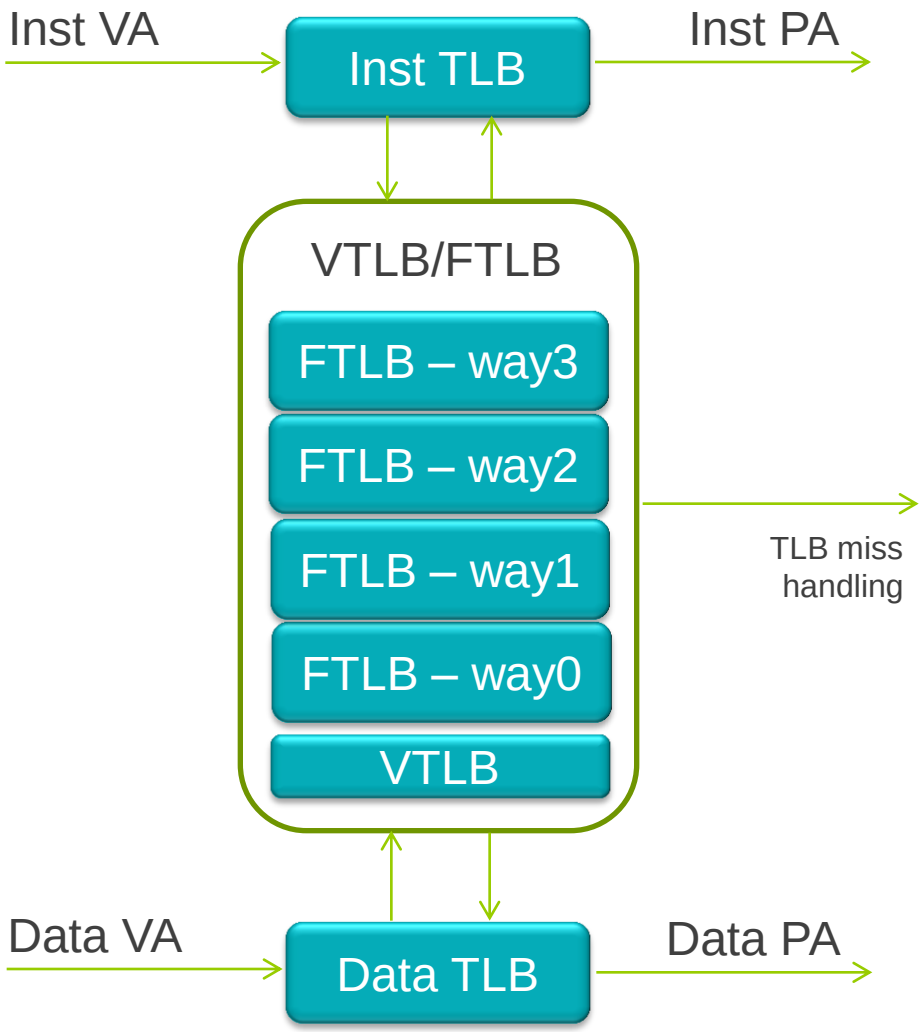
© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Memory Subsytem

- **Designed for large modern workloads**
 - Enhanced Virtual Addressing (EVA) allows efficient access > 3GB
 - Via programmable segments and new kernel load-store instructions
- **LSU**
 - Out-of-Order operation: loads/stores can (with some restrictions) overtake each other
 - Important for performance
 - And efficiency (maximizes utilization of single load-store pipe)
 - But requires:
 - Excellent memory disambiguation and “RAW” hazard avoidance
 - Overeager Load Predictor accessed before insertion into scheduler
 - LSU CAMs detect failure to forward from store buffer and trains predictor
 - Mark a specific load as overeager
 - Predictor forces marked loads to be uneager
 - Scheduler holds overeager loads until all older STA and STD have issued
 - Enforce MIPS’ **weakly-ordered** memory consistency model
 - Store merging, lightweight and heavyweight SYNCs, cache-ops
 - FP stores can graduate even before receiving store data from FPU
- **BIU**
 - Write-combining and bonding to support streaming writes

© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Memory Management



- **MIPS dual-entry scheme in TLBs**
 - Two VAs differing by 1 address bit share CAM/index portion of entry
 - Separate PA for each of the two VAs
- **Instruction and Data TLBs**
 - Holds 16KB or 4KB pages or sub-pages from VTLB/FTLB
 - 16 entry Instruction TLB
 - 32 dual entry Data TLB
 - Fast adder-comparator logic
- **Variable page size TLB (VTLB)**
 - 64 dual entries, fully associative
 - Holds pages from 4KB – 256MB
- **Fixed page size TLB (FTLB)**
 - 512 dual entries, 4-way assoc
 - Holds either 16KB or 4KB pages
 - Optional at build and runtime
 - SRAM-based implementation

© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Floating Point

- **Brand new high-speed design**
 - Can run 1:1 with proAptiv up to top achievable core frequency
 - Native double-precision datapath
 - FMAC-based pipeline with early and late bypass for FADD/FMUL
 - 4-cycle FADD, 4-cycle FMUL, 7-cycle FMAC
 - Low latency and high throughput for long ops like div/sqrt/recip/rsqrt
 - Functional iterative algorithms and lookup tables compared to bitwise SRT
 - Can run independent instructions under a long op, including other long ops
- **Coprocessor style FPU**
 - Has its own decoupled pipeline, regfile and load/store interface buffers
 - Non-stalling design using shelving buffers to reduce power, improve perf
 - Lower power than PRF-style renaming, given flop-based implementation
- **Formal verification**
 - Against a precise IEEE-compliant mathematical model

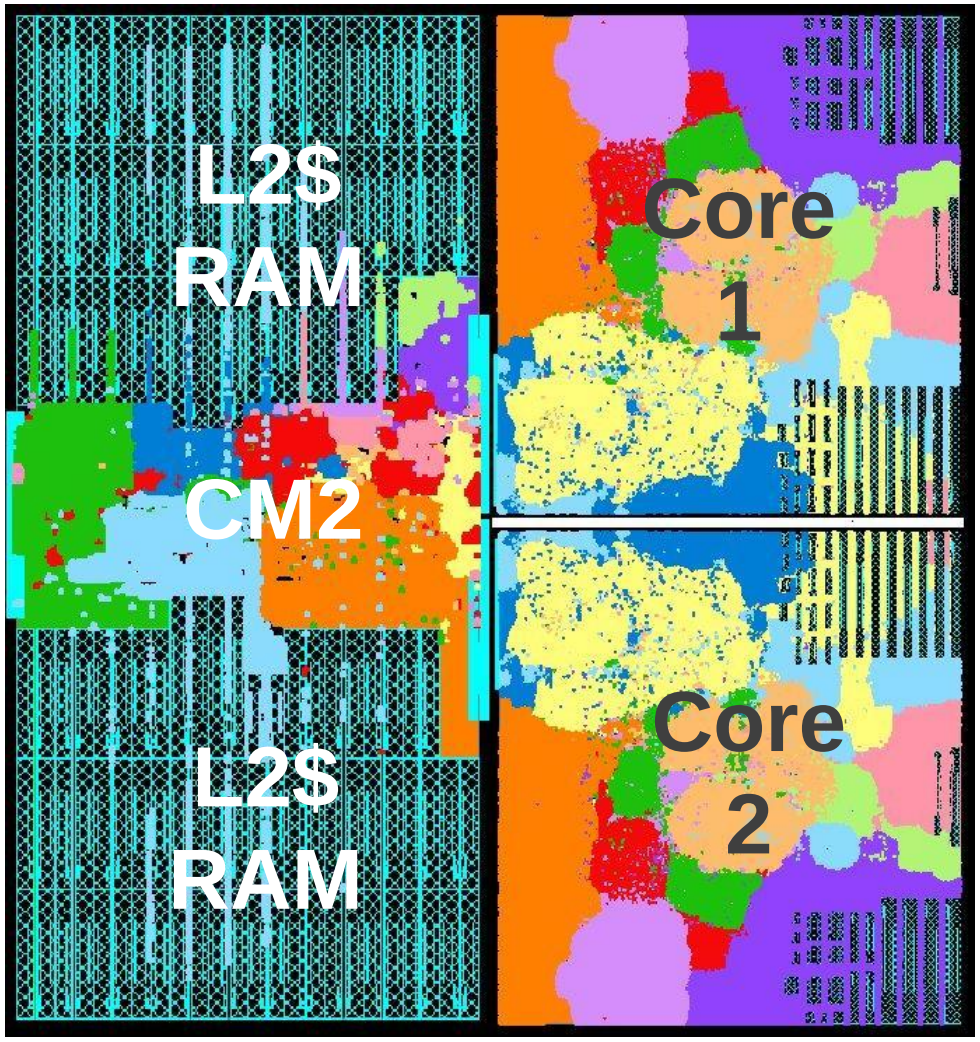
© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv L2 Cache Controller

- Accompanies both proAptiv and interAptiv cores
- 256KB to 8MB shared across 1 to 6 cores
- 8-way associative
- Selectable 32 or 64B line size
- 256-bit internal datapaths and buffers
- Up to 256-bit interface to system interconnect
- Optional wait states on tag, data or control RAMs
 - Accommodates slow memories, due to:
 - Large size or high-frequency operation
 - HD bitcells, pipelined RAMs, low-voltage operation
- Optional ECC on all RAMs
 - Adds one pipestage
- L2 storage non-inclusive to L1
- Critical-word first; can interleave responses to multiple cores

© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Dual-Core Floorplan

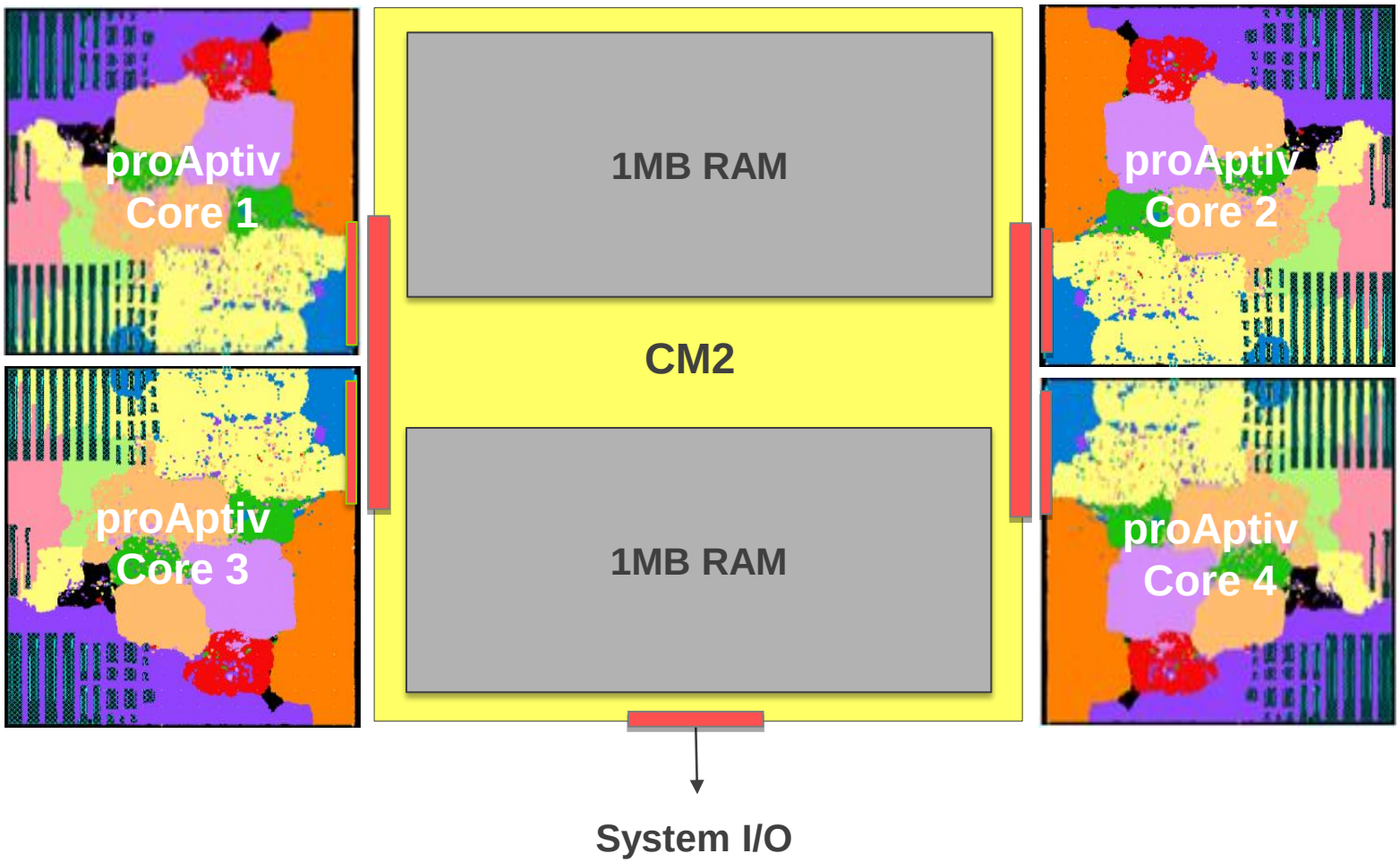


Configuration:

- **Per base core**
 - FPU
 - 32KB/32KB I/D L1$s
 - TLB
 - I and D TLBs
 - 128 entry VTLB
 - 1024 entry FTLB
 - PDtrace
- **Cluster level**
 - Dual core coherence
 - 1MB L2$ with ECC
 - PDtrace aggregator
 - 64-Interrupt Controller
 - HW IO coherence
 - Cluster power controller
 - Probe interface block

© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Quad-Core Floorplan



© 2012 MIPS Technologies, Inc. All rights reserved.

proAptiv Summary

- ***Fast***
 - *4.5 EEMBC CoreMark/MHz*
 - Highest single-threaded score published for any licensable CPU*
 - 75% over prior MIPS 1074K core
 - Operating frequency > 1GHz worst-case, >> 2GHz typical at 40nm
- ***Slim***
 - Highest CoreMark/mm^2 for any licensable CPU*
 - Dual core area ~ 1MB L2 cache
- ***Cool***
 - Highest CoreMark/mW for any licensable CPU*
 - Sub half-watt power at 40nm

- ***Efficient performance* on a fully-synthesizable core**

* CoreMark/MHz derived from publicly available and published scores at http://www.coremark.org
Area and power efficiencies based on MIPS internal and competitive estimates

© 2012 MIPS Technologies, Inc. All rights reserved.

Thank You!

Questions?



At the core of the user experience®

MIPS, MIPS I, MIPS II, MIPS III, MIPS IV, MIPS V, MIPSr3, MIPS32, MIPS64, microMIPS32, microMIPS64, MIPS-3D, MIPS16, MIPS16e, MIPS-Based, MIPSsim, MIPSpro, MIPS Technologies logo, MIPS-VERIFIED, MIPS-VERIFIED logo, 4K, 4Kc, 4Km, 4Kp, 4KE, 4KEc, 4KEm, 4KEp, 4KS, 4KSc, 4KSd, M4K, M14K, 5K, 5Kc, 5Kf, 24K, 24Kc, 24Kf, 24KE, 24KEc, 24KEf, 34K, 34Kc, 34Kf, 74K, 74Kc, 74Kf, 1004K, 1004Kc, 1004Kf, 1074K, 1074Kc, 1074Kf, R3000, R4000, R5000, Aptiv, ASMACRO, Atlas, "At the core of the user experience.", BusBridge, Bus Navigator, CLAM, CorExtend, CoreFPGA, CoreLV, EC, FPGA View, FS2, FS2 FIRST SILICON SOLUTIONS logo, FS2 NAVIGATOR, HyperDebug, HyperJTAG, IASim, interAptiv, JALGO, Logic Navigator, Malta, MDMX, MED, MGB, microAptiv, microMIPS, OCI, PDtrace, the Pipeline, proAptiv, Pro Series, SEAD, SEAD-2, SmartMIPS, SOC-it, System Navigator, and YAMON are trademarks or registered trademarks of MIPS Technologies, Inc. in the United States and other countries.

© 2012 MIPS Technologies, Inc. All rights reserved.

Swizzle Switch: A Self-Arbitrating High-Radix Crossbar for NoC Systems



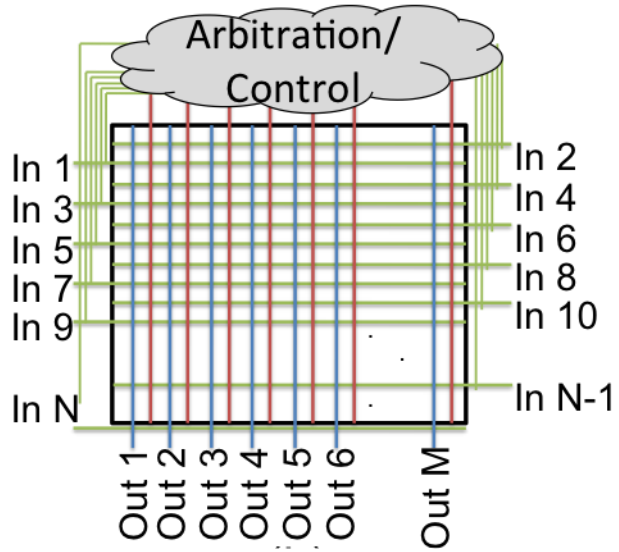
Ronald Dreslinski, Korey Sewell, Thomas Manville, Sudhir Satpathy, Nathaniel Pinckney, Geoff Blake, Michael Cieslak, Reetuparna Das, Thomas Wenisch, Dennis Sylvester, David Blaauw, and Trevor Mudge

University of Michigan

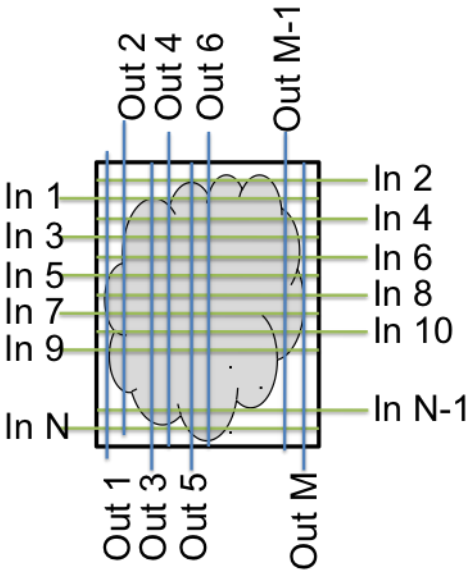
Outline

- Swizzle Switch—Circuit & Microarchitecture
 - Overview
 - Arbitration
 - Prototype
- Swizzle Switch—Cache Coherent Manycore Interconnect
 - Motivation & Existing Interconnects
 - Swizzle Switch Interconnect
 - Evaluation

Swizzle Switch



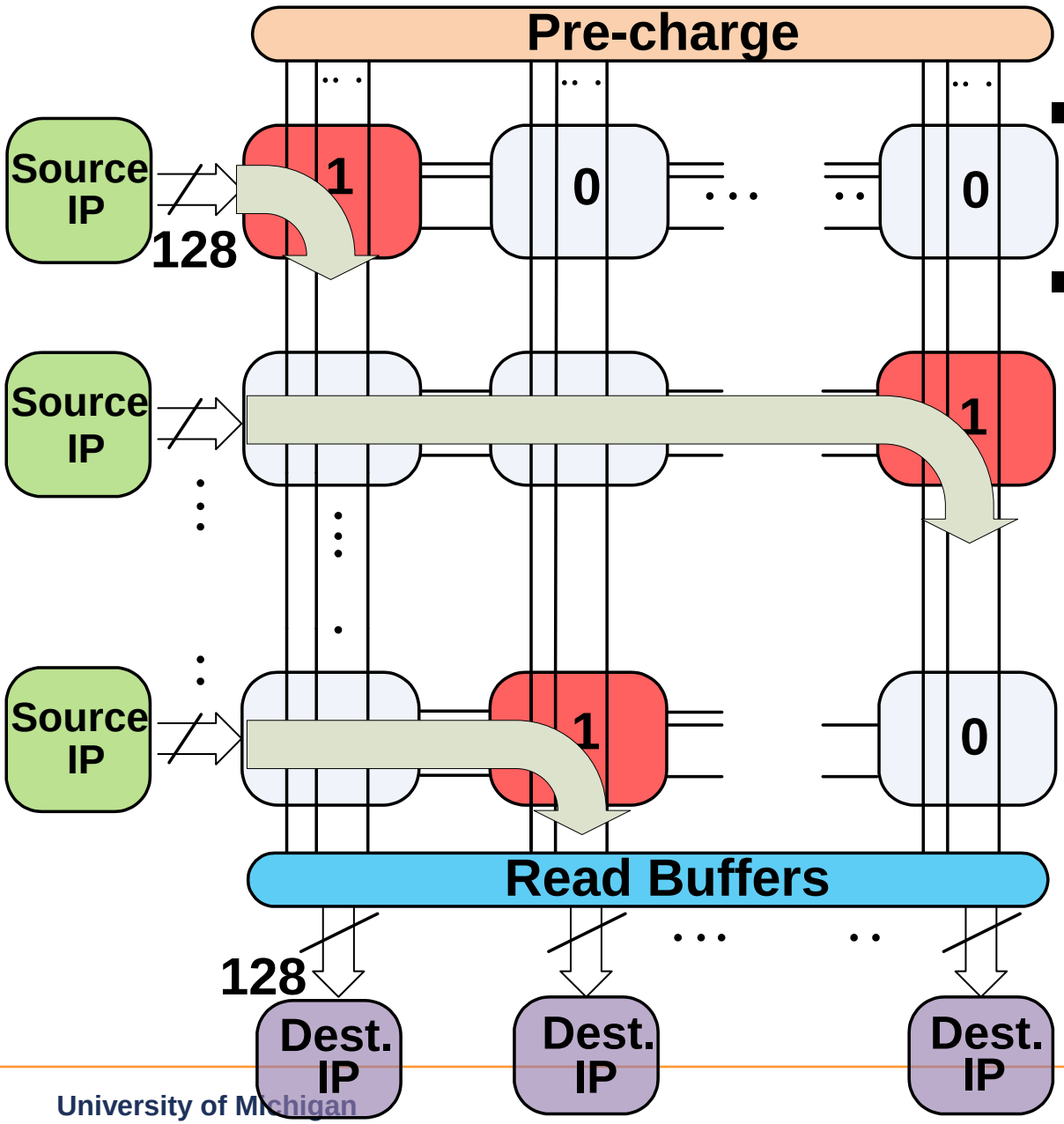
Conventional Matrix Crossbar



Swizzle Switch

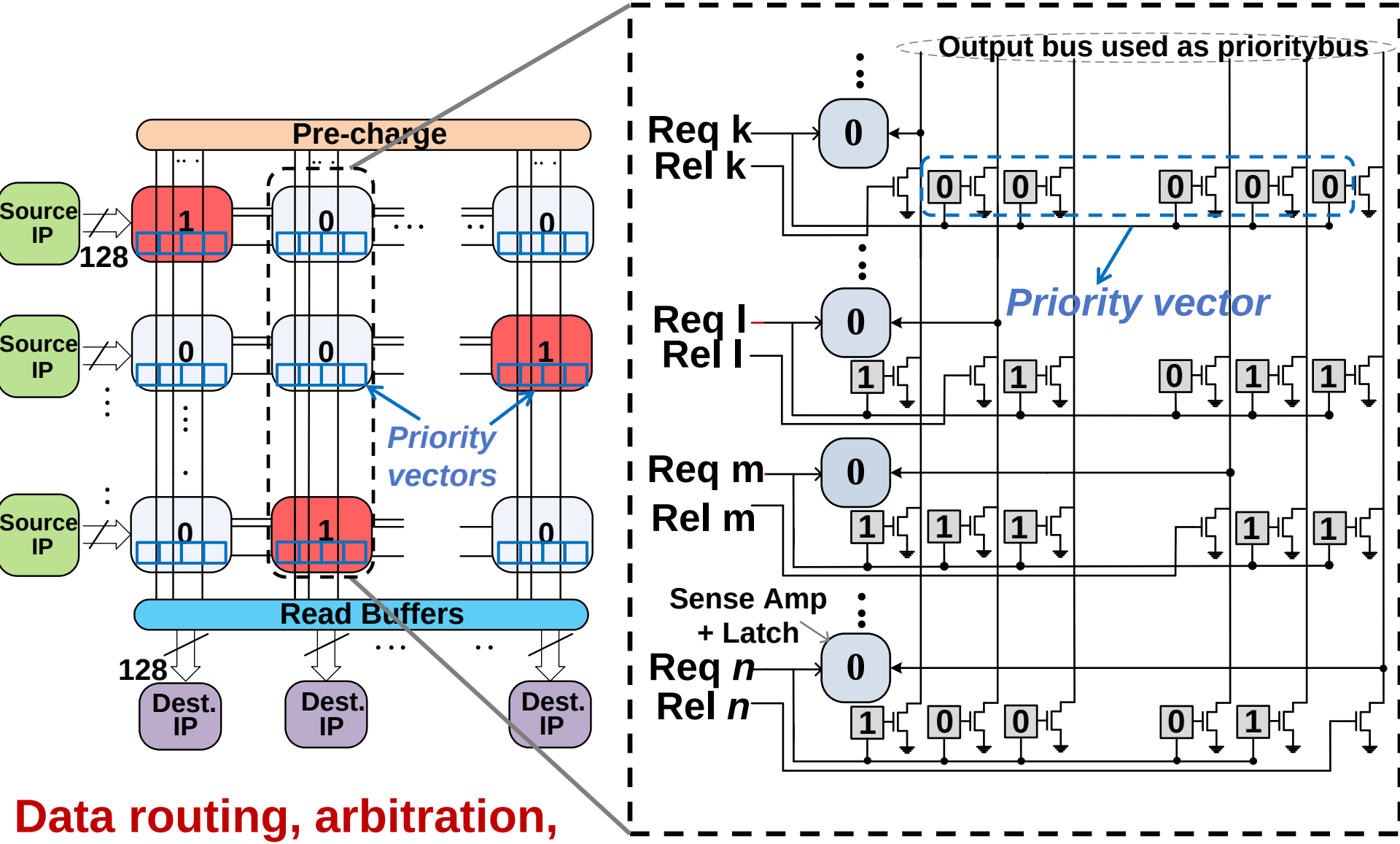
- Embeds arbitration within crossbar—single cycle arbitration
- Re-use input/output data buses for arbitration
- SRAM-like layout with priority bits at cross-points
- Low-power optimizations
- Excellent scalability

Data Routing



- Multicast & Broadcast
- Bitlines discharged if
 - Data = “1”
 - Crosspoint = “1”

Swizzle Switch Architecture

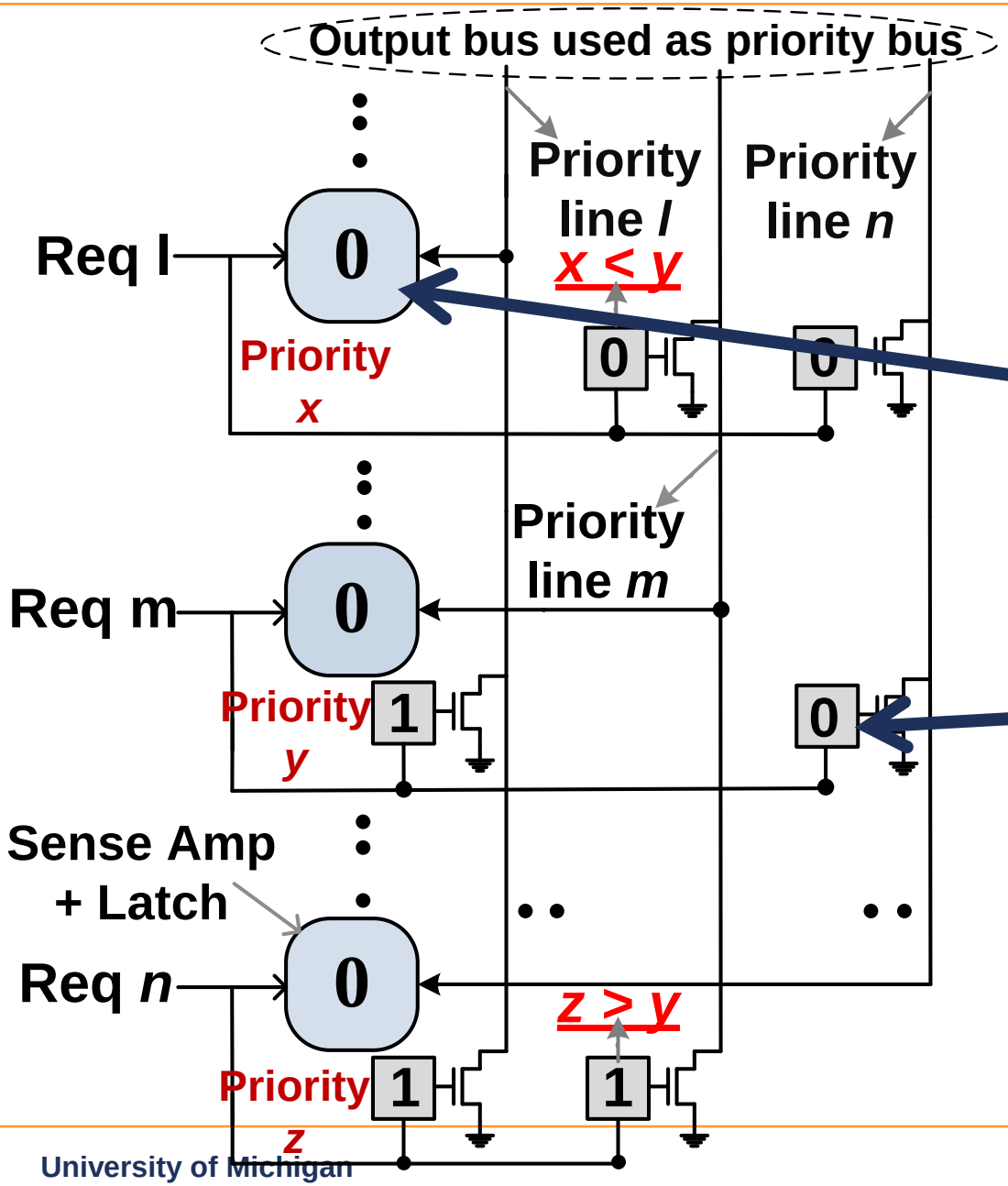


Data routing, arbitration,
And priority update control embedded within crosspoints

Outline

- Swizzle Switch—Circuit & Microarchitecture
 - Overview
 - Arbitration
 - Prototype
- Swizzle Switch—Cache Coherent Manycore Interconnect
 - Motivation & Existing Interconnects
 - Swizzle Switch Interconnect
 - Evaluation

Inhibit Based Arbitration



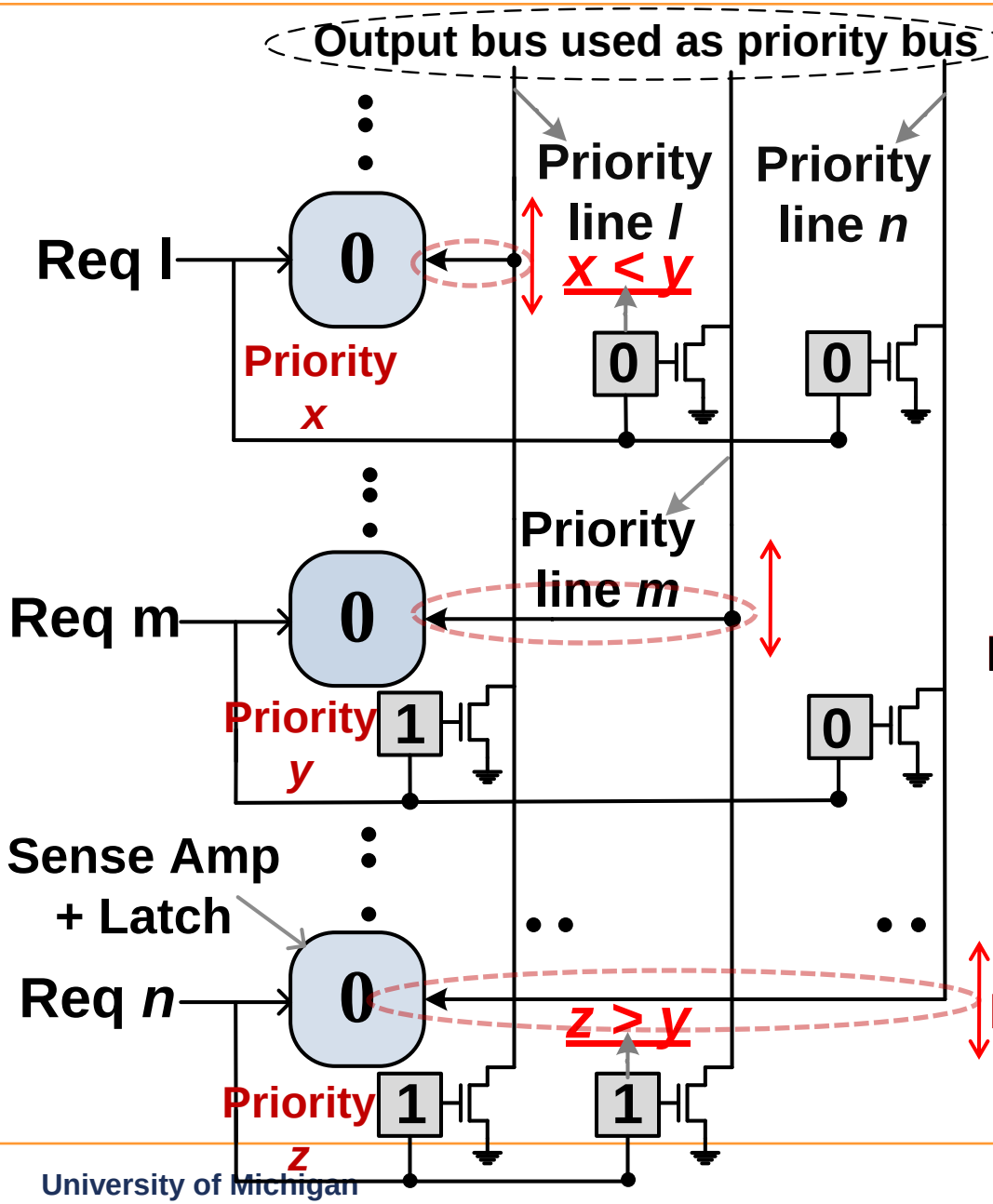
This diagram is a single column in the Swizzle-Switch (output), each output arbitrates/transfers data ***independently***

Each Crosspoint has a sense amp/latch to indicate connectivity. Each input samples a unique bit of the output bus to determine if it has been granted the channel

Priority vectors are stored and when a request is issued they discharge bits along the output columns to ***INHIBIT*** lower priority requests

Finally, the priority vectors are updated when the data transfer completes.

Least Recently Granted(LRG)

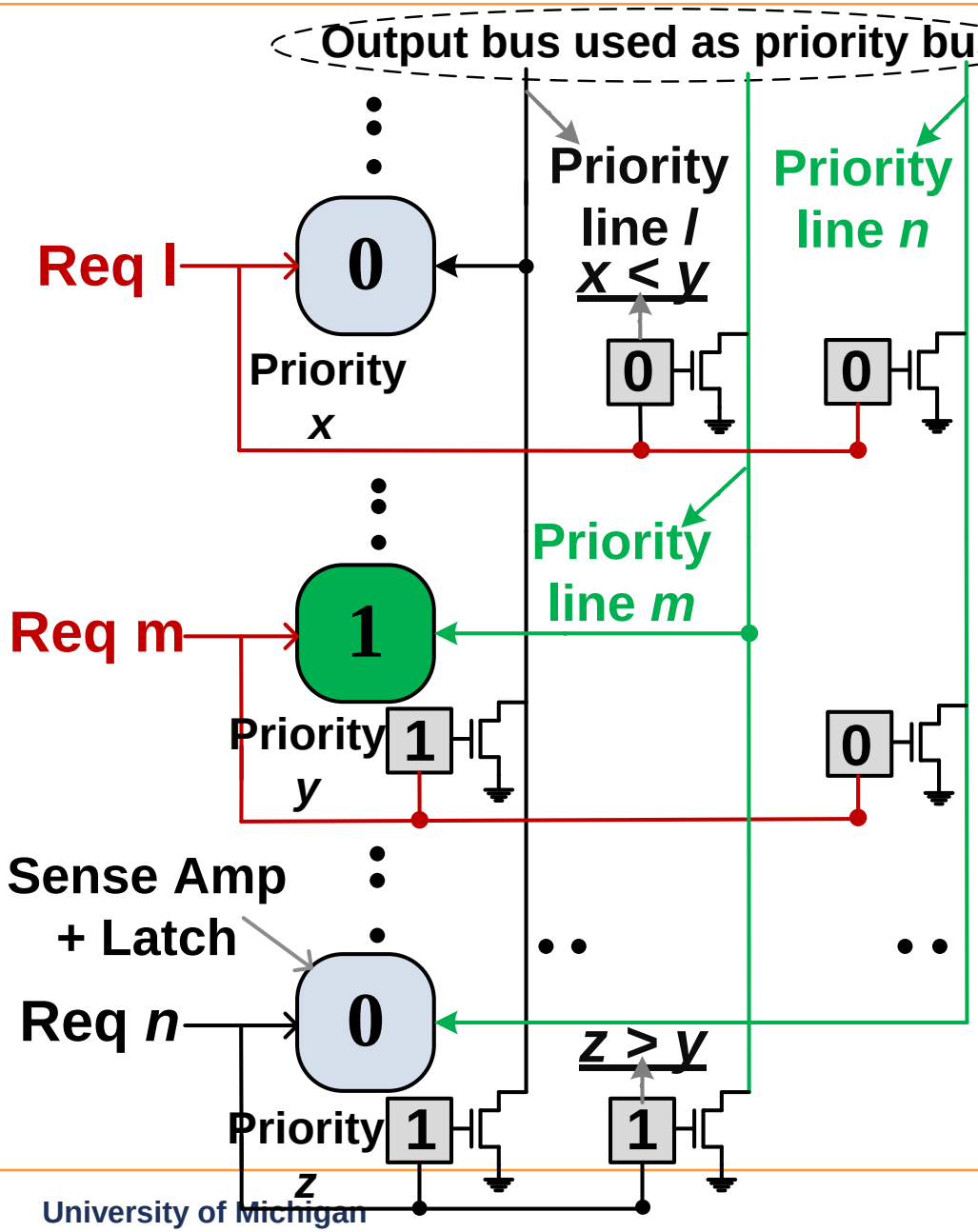


LOWEST Priority
Discharges NO Priority Lines

INTERMEDIATE Priority
Discharges SOME Priority Lines

HIGHEST Priority
Discharges ALL Priority Lines

Least Recently Granted(LRG)



Example Arbitration:

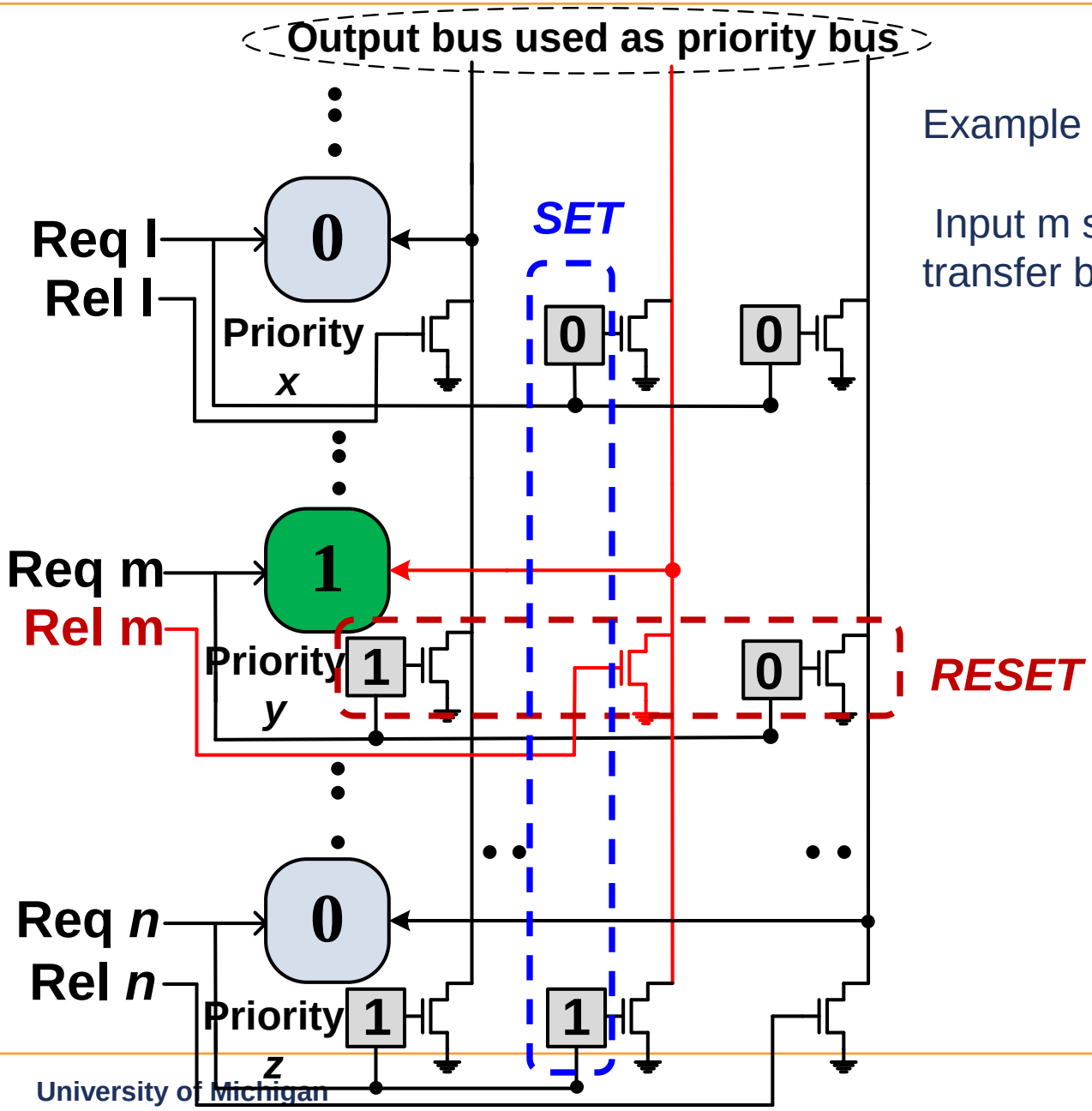
(1) ***Req l*** and ***Req m*** Request the bus (red lines)

(2) ***Req m*** discharges Priority line *l*, priority lines *m* and *n* remain charged (green lines)

(3) ***Req l*** senses Priority line *l* and is inhibited (not granted)***, Req m*** senses Priority line m and is not inhibited

(4) The crosspoint records the connectivity at ***input m***

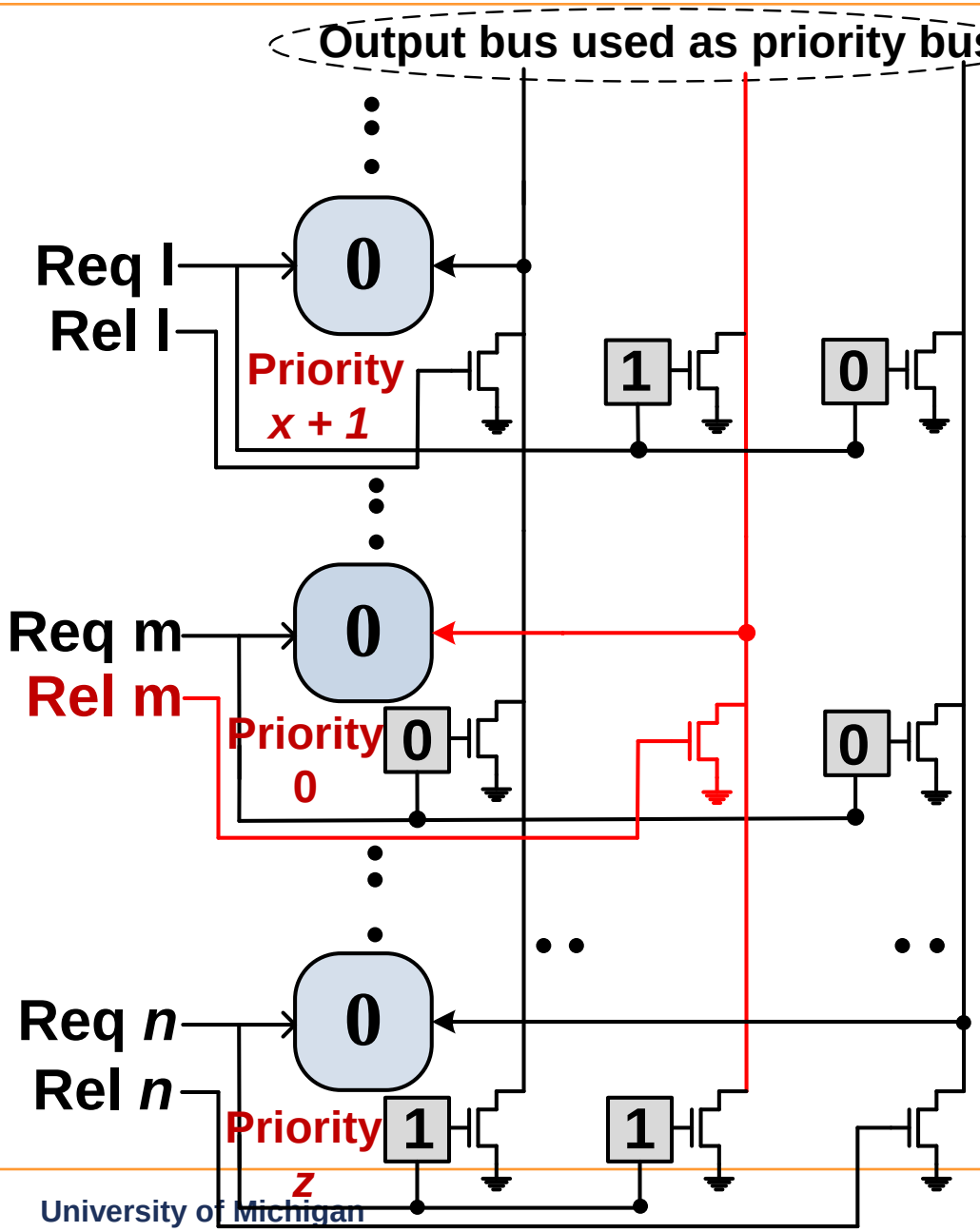
Least Recently Granted(LRG)



Example Priority Update:

Input m signals it is done with data transfer by asserting ***Rel m***

Least Recently Granted(LRG)

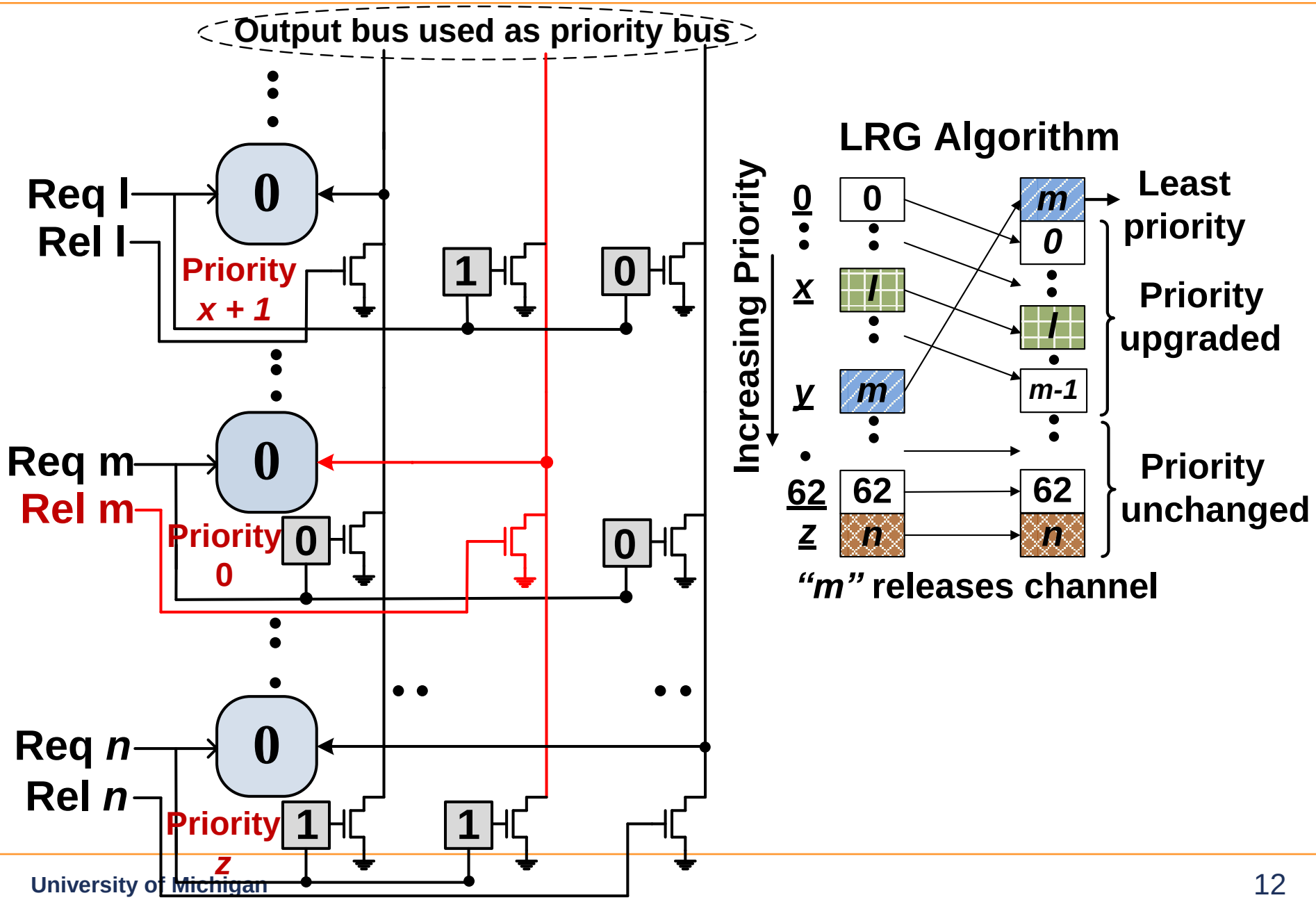


INTERMEDIATE Priority
Discharges SOME Priority Lines

LOWEST Priority
Discharges NO Priority Lines

HIGHEST Priority
Discharges ALL Priority Lines

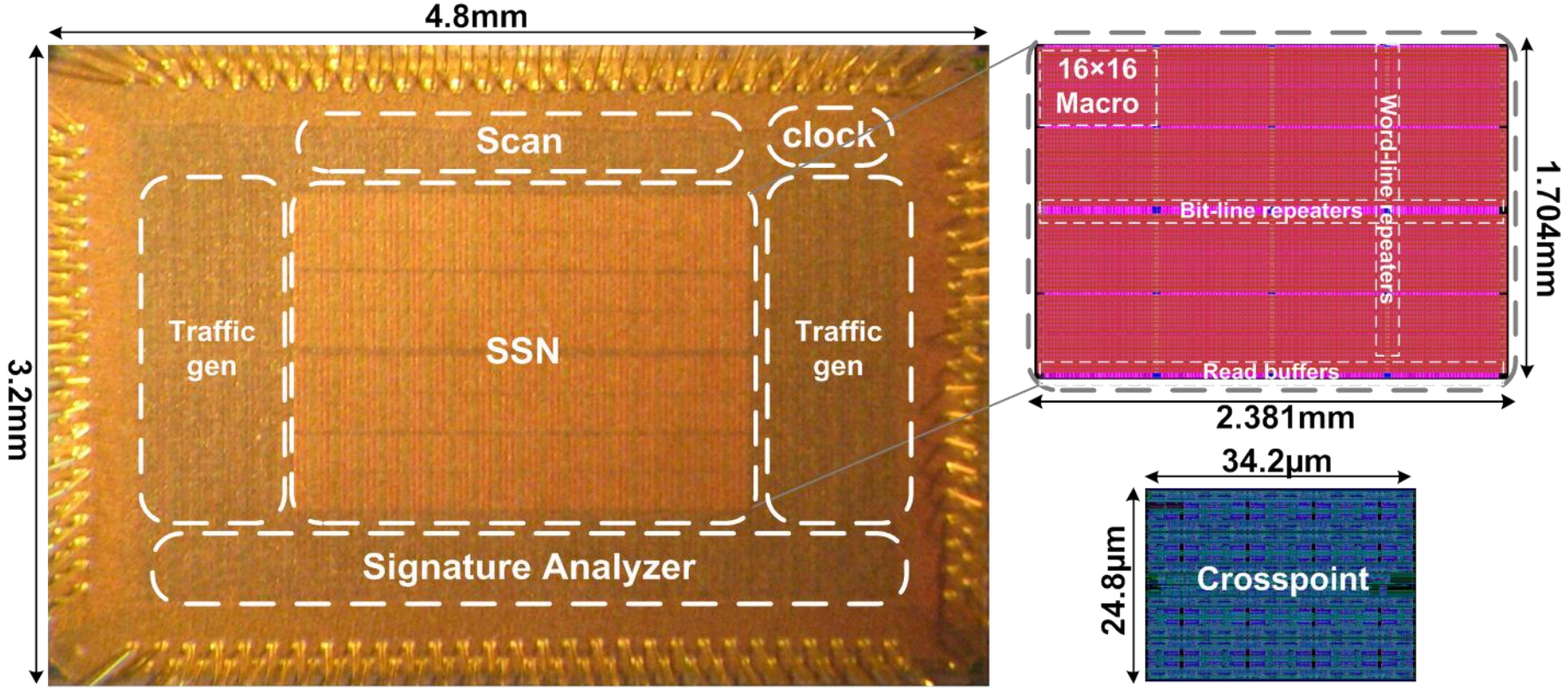
Least Recently Granted(LRG)



Outline

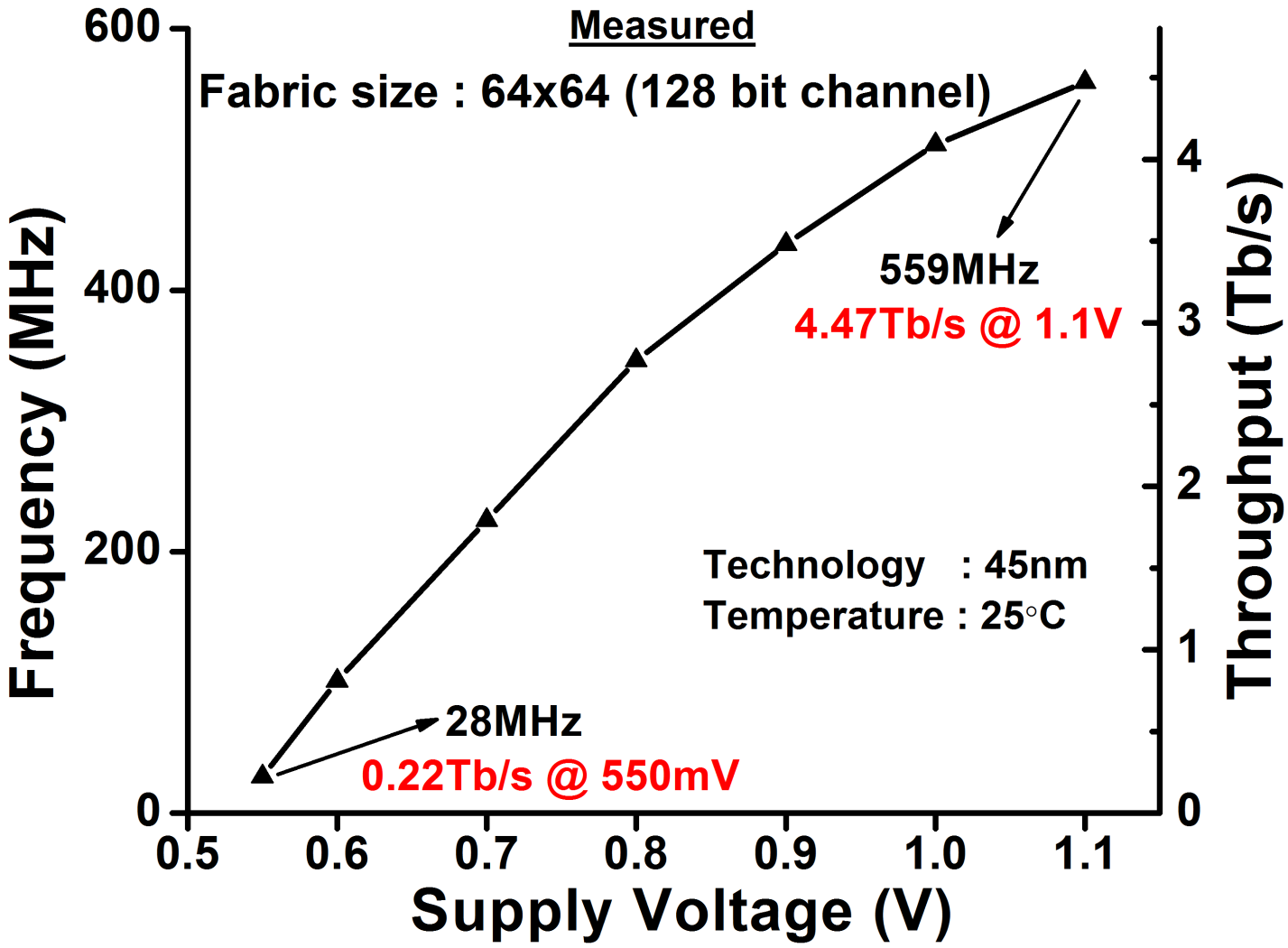
- Swizzle Switch—Circuit & Microarchitecture
 - Overview
 - Arbitration
 - Prototype
- Swizzle Switch—Cache Coherent Manycore Interconnect
 - Motivation & Existing Interconnects
 - Swizzle Switch Interconnect
 - Evaluation

64x64 Prototype

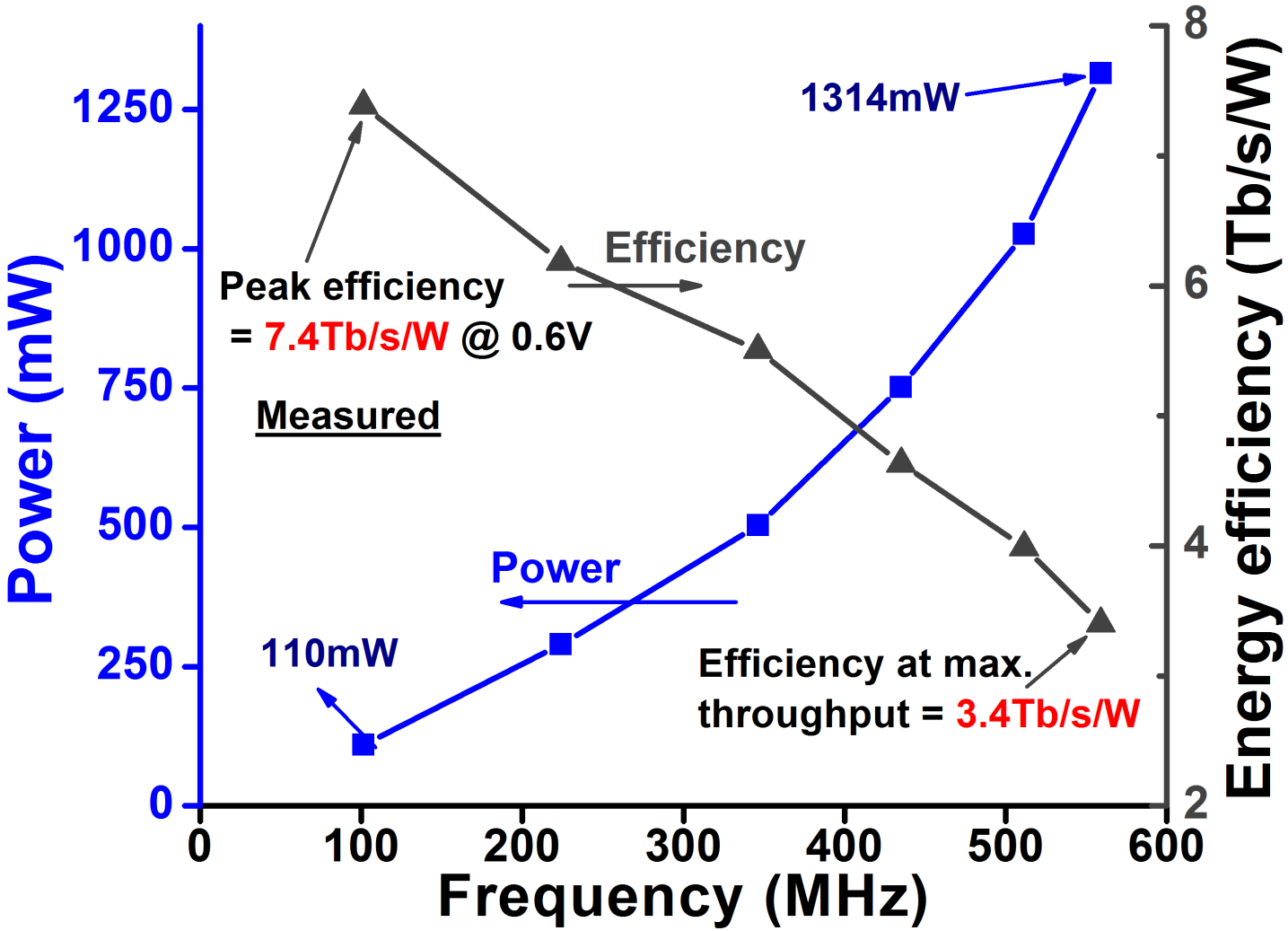


Process	45nm SOI CMOS 12metal interconnect
Die area	15.6mm^2
Fabric area, Transistor count, # Data wires	4.06mm^2, 6.95M, 8192
Throughput, Frequency	4.47Tb/s @ 1.1V, 559MHz, 25°C
Energy Efficiency at peak throughput	3.4Tb/s/W
Peak energy efficiency	7.4Tb/s/W @ 0.6V

Measurement Results



Measurement Results



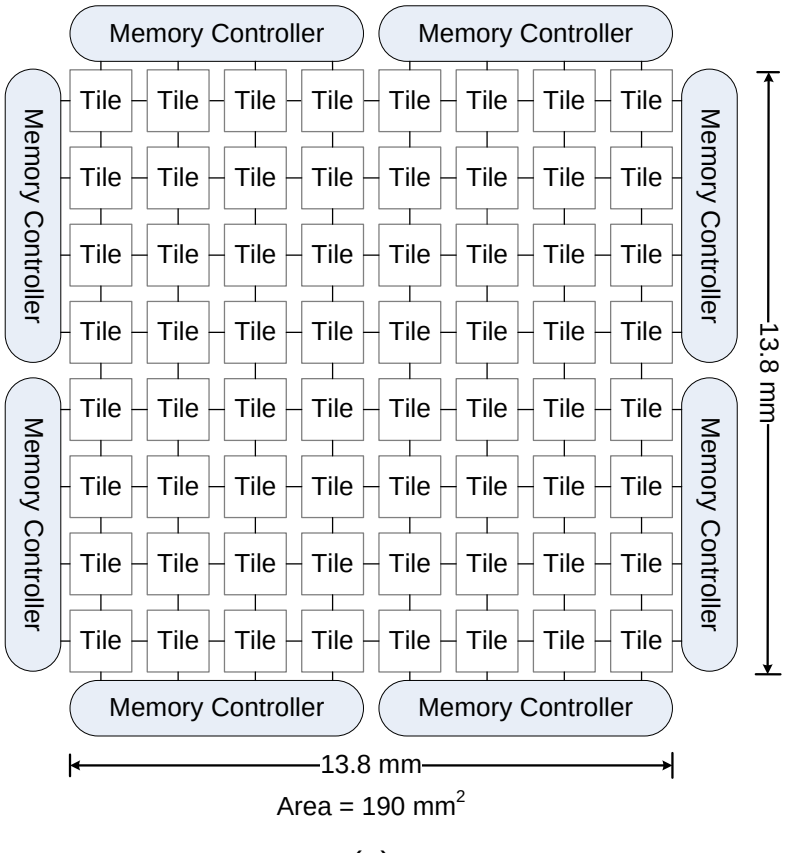
Outline

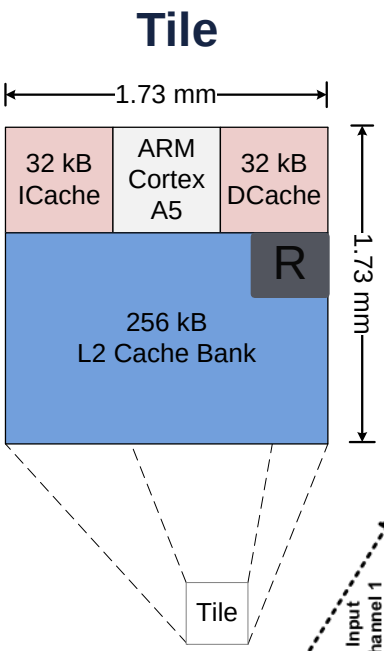
- Swizzle Switch—Circuit & Microarchitecture
 - Overview
 - Arbitration
 - Prototype
- Swizzle Switch—Cache Coherent Manycore Interconnect
 - Motivation & Existing Interconnects
 - Swizzle Switch Interconnect
 - Evaluation

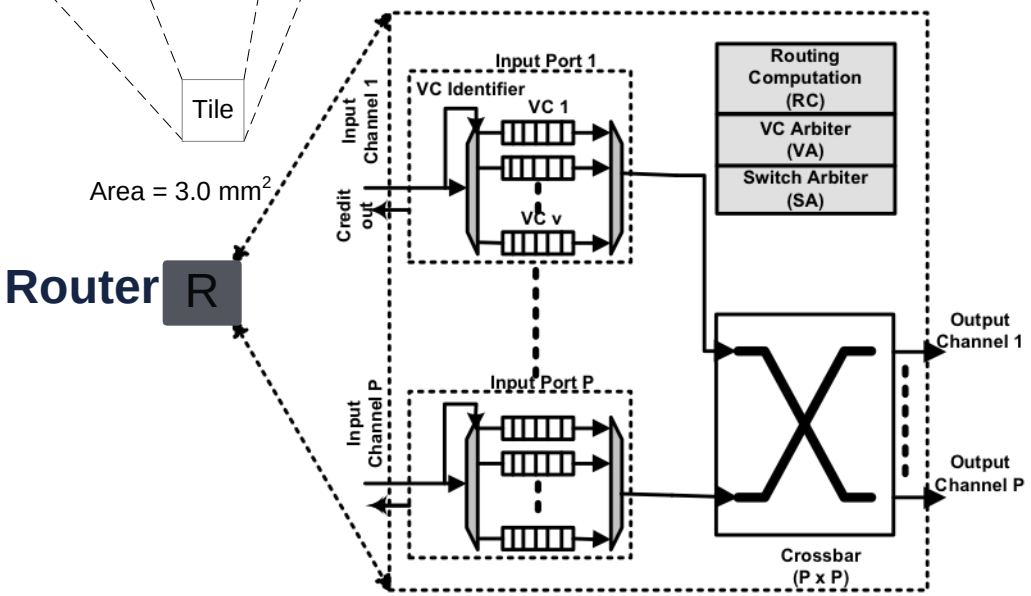
Scaling Interconnect for Many-Cores

- Existing interconnects—Buses, Crossbars, Rings
 - Limited to ~16 cores
- Other's Interconnect proposals for Many-Cores
 - Packet-switched, multi-hop, network-on-chip (NoC)
 - Grid of routers—meshes, tori and flattened butterfly
- **Our Proposal**
 - **Swizzle Switch Networks**
 - Flat single-stage, one-hop, crossbar++ interconnect

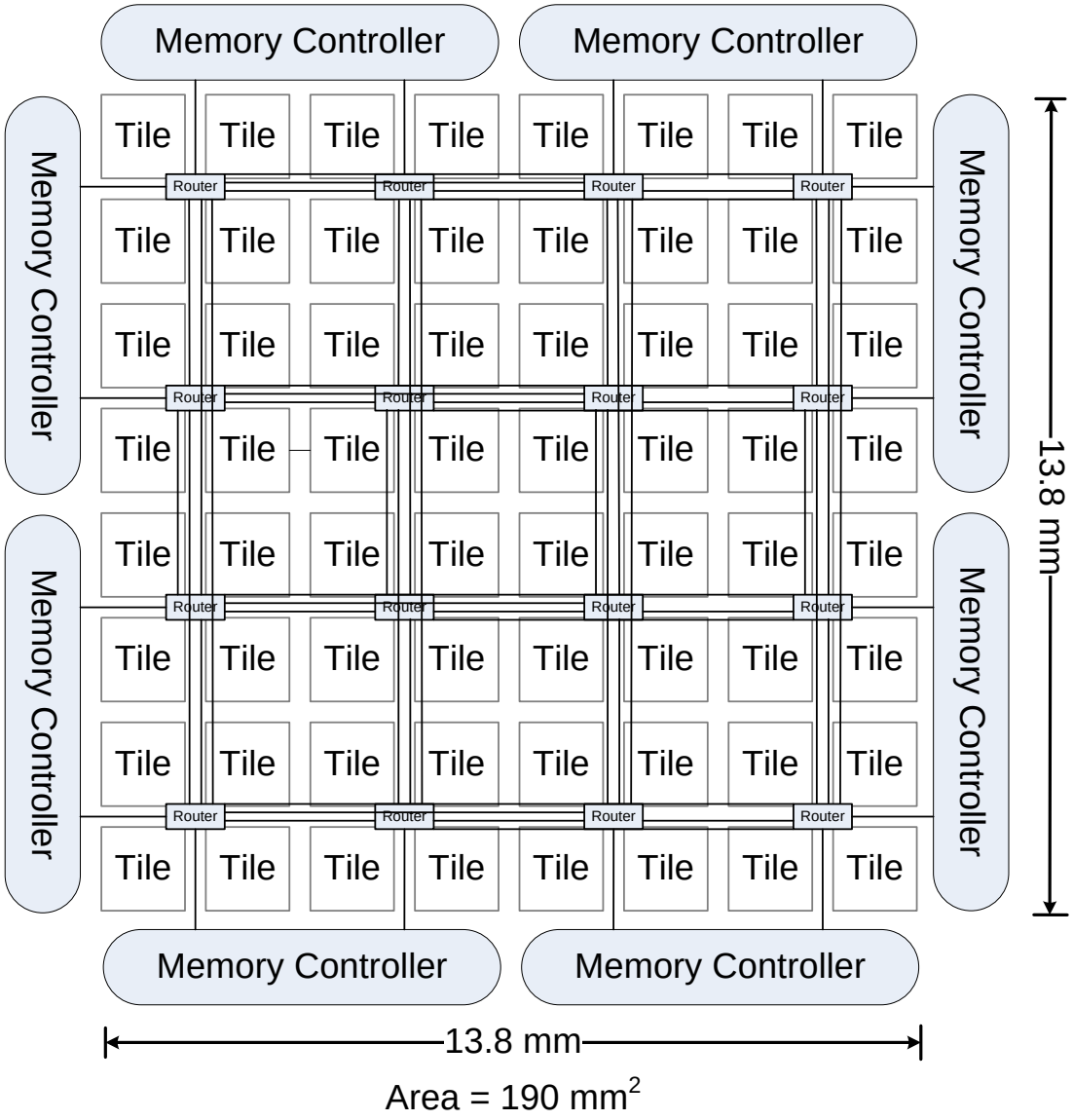
Mesh Network-on-Chip







Flattened Butterfly Network-on-Chip

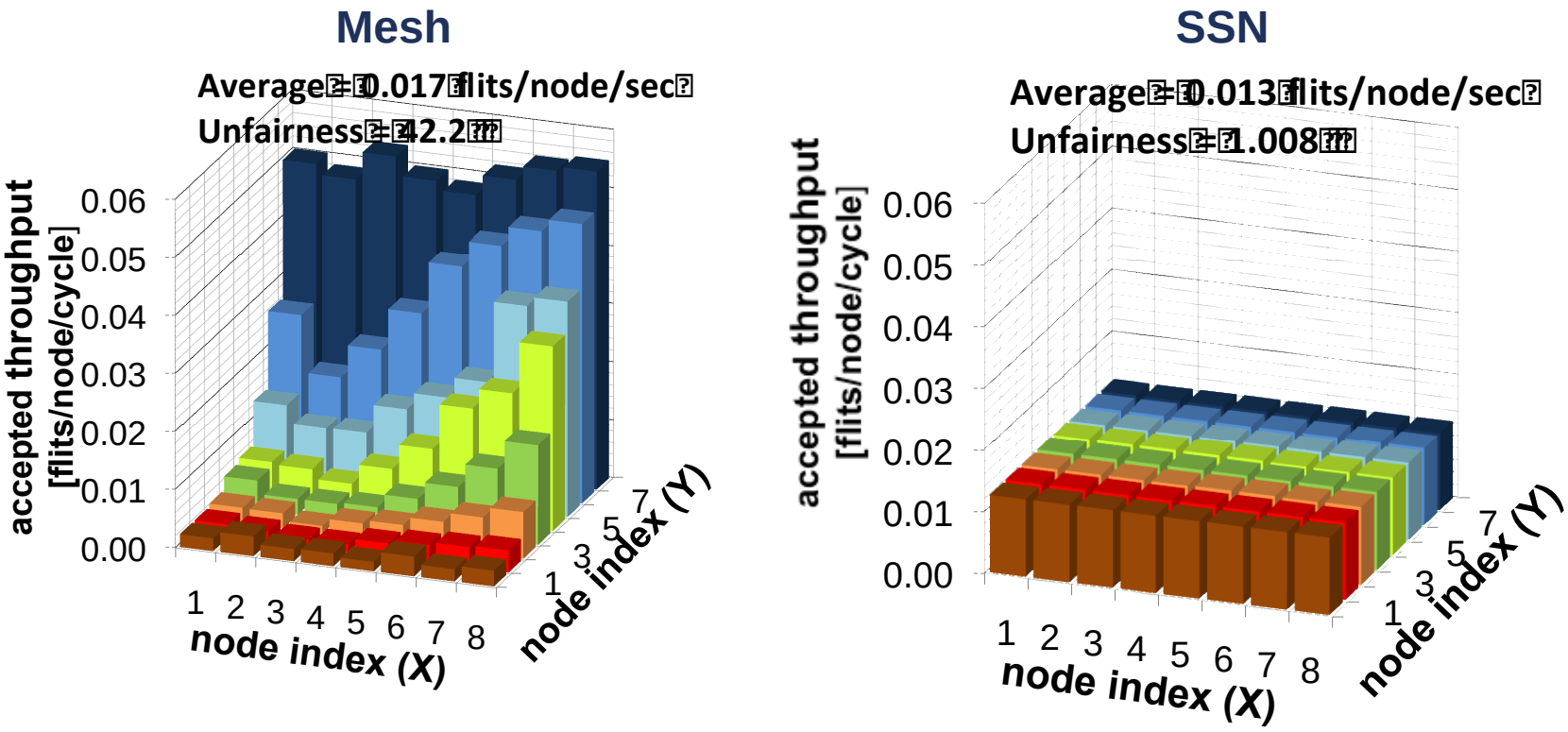


Area = 190 mm^2

Motivating Swizzle Switch Networks

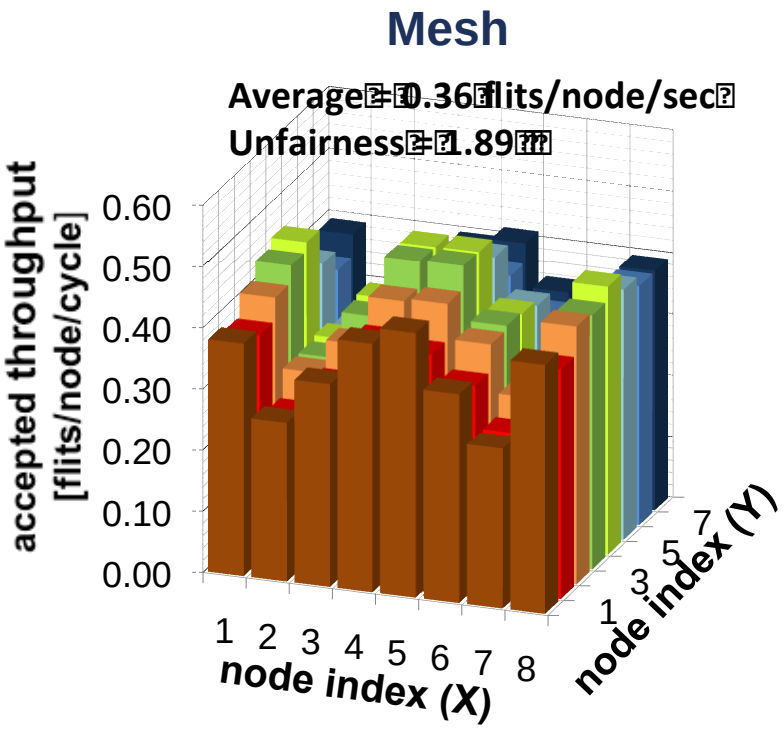
- Uniform access latency
 - Ease of programming, data placement, thread placement,...
- Low Power
- Simplicity
 - Packet-switched NoCs need routing, congestion management, flow control, wormhole switching,...

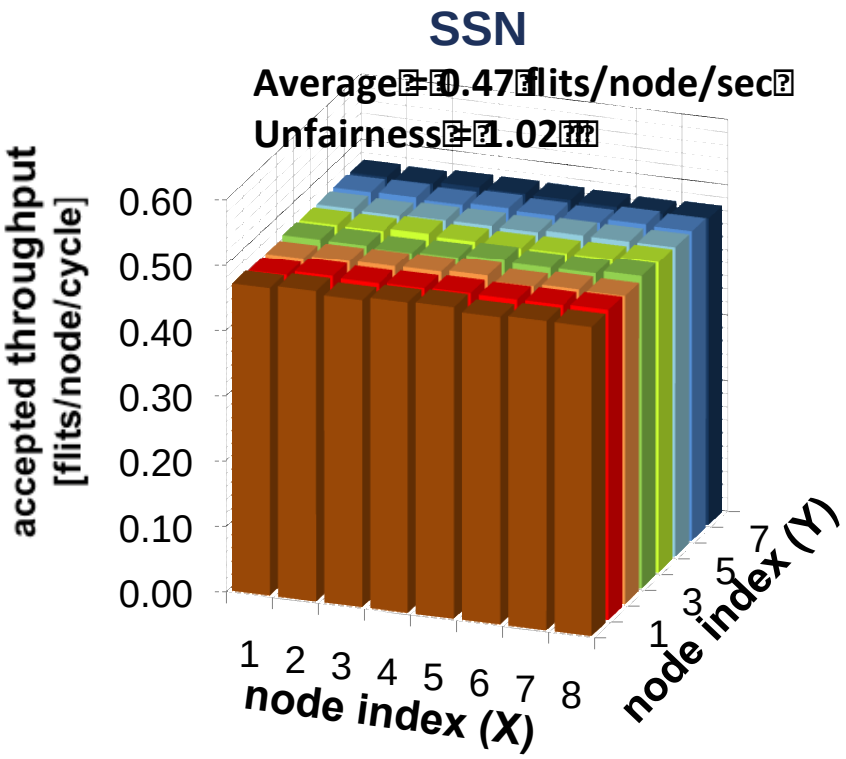
Motivating Swizzle Switch Networks



- Unfairness = $Node_{highest_throughput}$ / $Node_{lowest_throughput}$
- Hotspot Traffic = All nodes sending data to $node_{8,8}$
 - Under Hotspot traffic, the Crossbar has a slightly less throughput than the Mesh but is *40x more fair*.

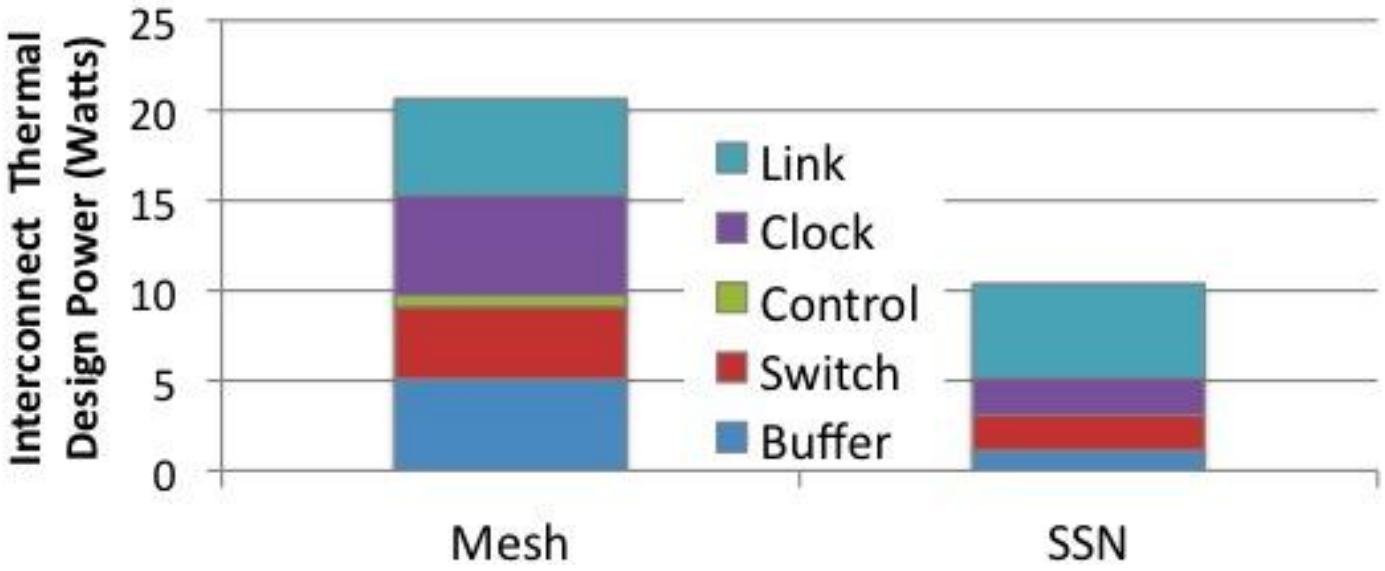
Motivating Swizzle Switch Networks





- In the Mesh, nodes closest to the center receive the highest throughput
- Under Uniform Random traffic, the Crossbar has more throughput than the Mesh and is *87% more fair*.

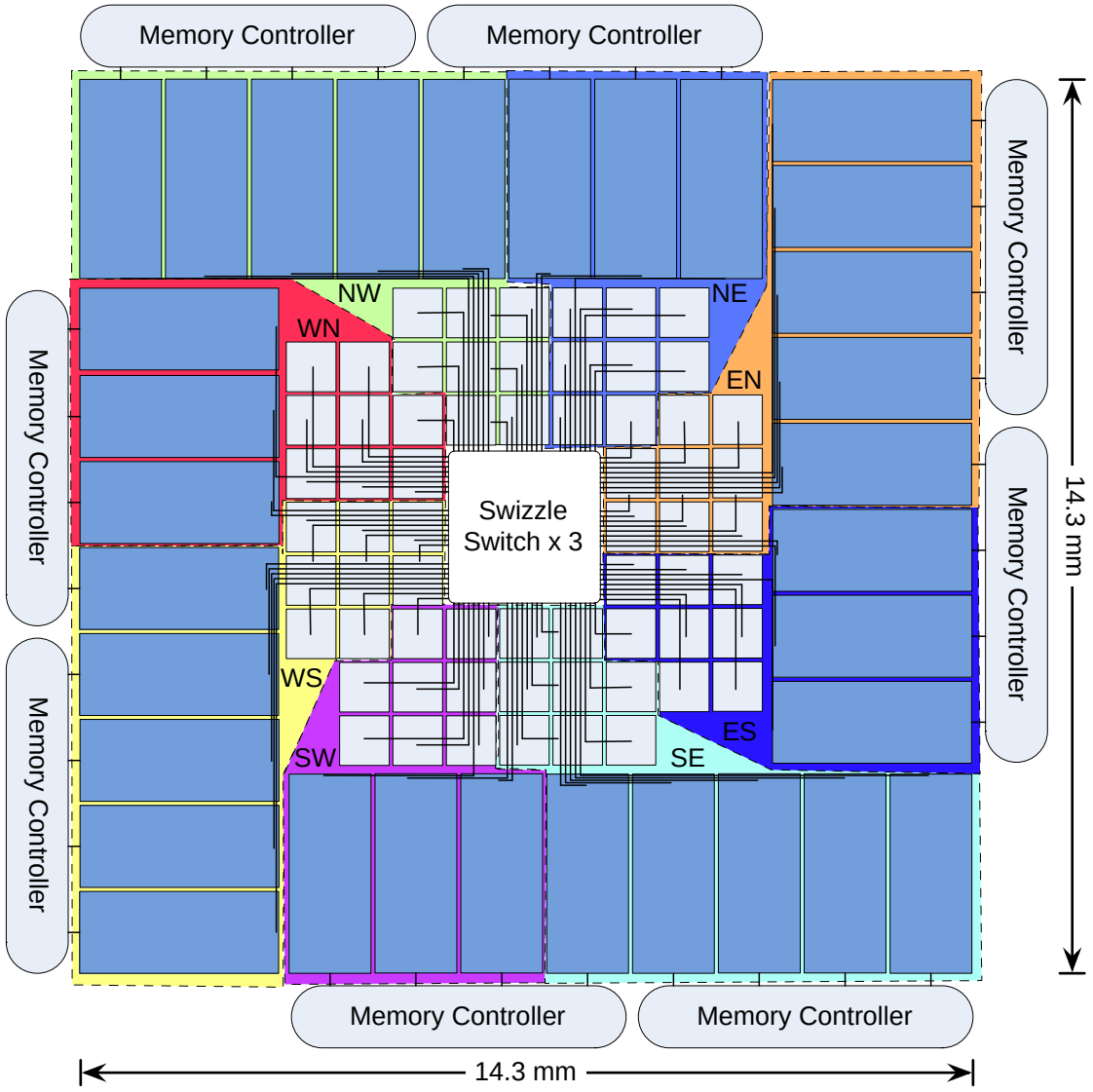
Motivating Swizzle Switch Networks



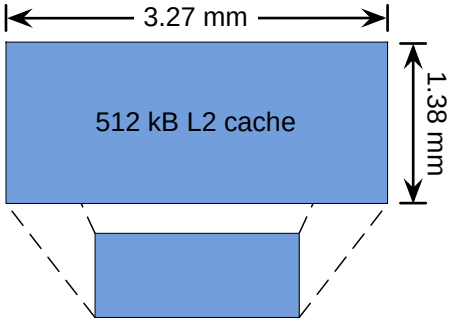
Outline

- Swizzle Switch—Circuit & Microarchitecture
 - Overview
 - Arbitration
 - Prototype
- Swizzle Switch—Cache Coherent Manycore Interconnect
 - Motivation & Existing Interconnects
 - Swizzle Switch Interconnect
 - Evaluation

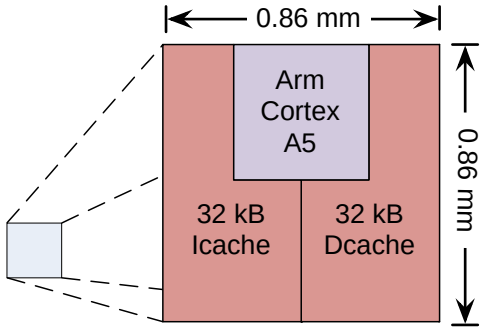
Top-Level Floorplan



Total Area = 204 mm^2



L2 Area = 4.50 mm^2



Core + L1 Area = .74 mm^2

Outline

- Swizzle Switch—Circuit & Microarchitecture
 - Overview
 - Arbitration
 - Prototype
- Swizzle Switch—Cache Coherent Manycore Interconnect
 - Motivation & Existing Interconnects
 - Swizzle Switch Interconnect
 - Evaluation

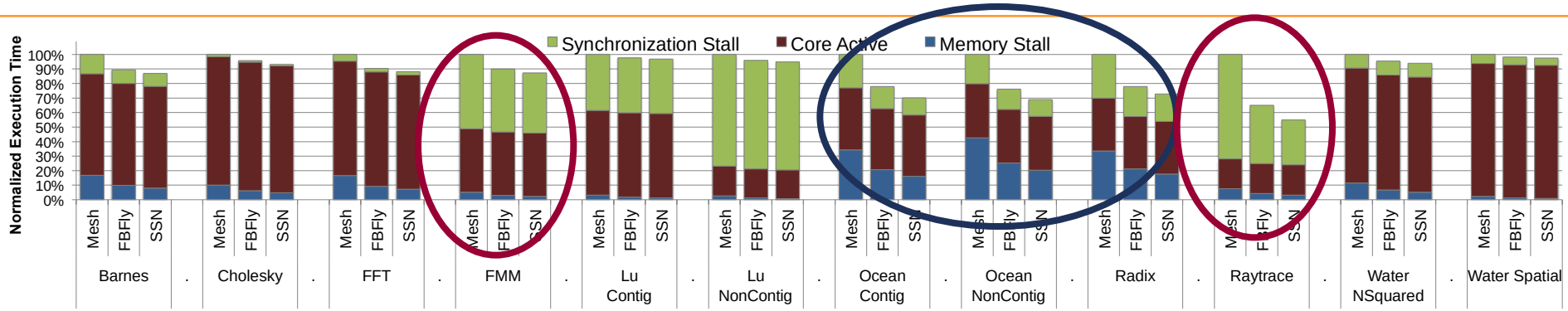
Evaluation

- Simulation Parameters

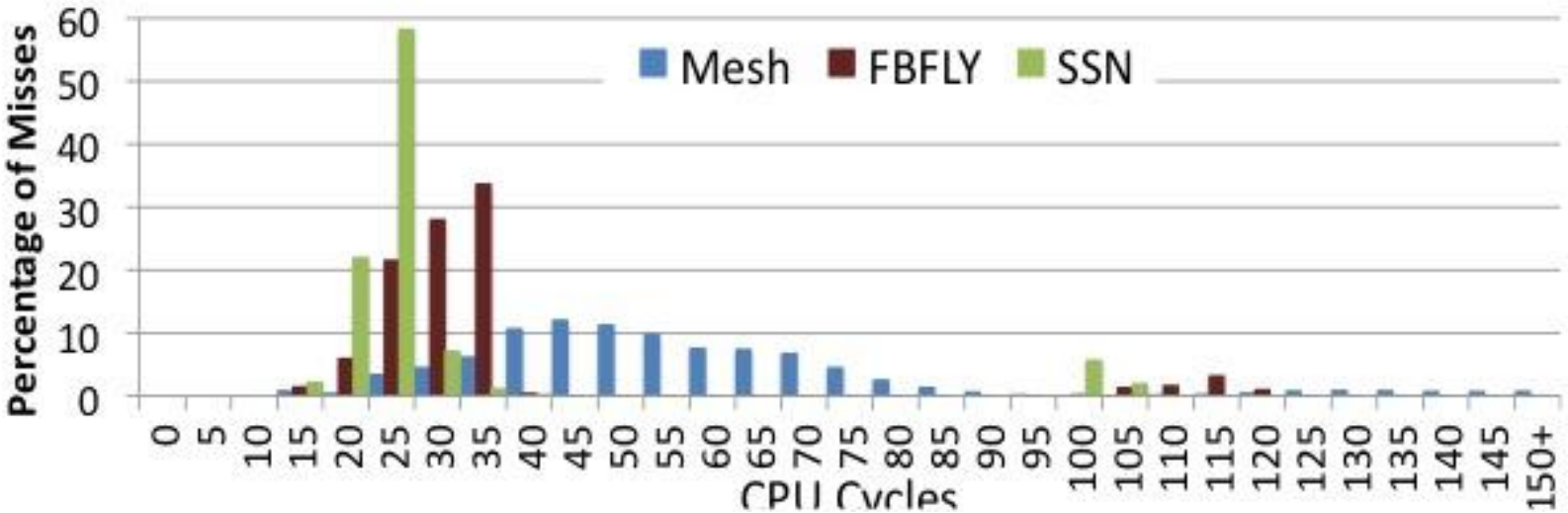
Feature	NoC (Mesh/FBFly)	SSN
Processors	64 in-order cores, 1 IPC, 1.5 GHz	
L1 Cache	32kB I/D Caches, 4-way associative, 64-byte line size, 1 cycle latency	
L2 Cache	Shared L2, 16 MB, 64-way banked, 8-way associative, 64-byte line size, 10 cycle latency	Shared L2, 16MB, 32-way banked, 16-way associative, 64-byte line size, 11 cycle latency
Interconnect	3.0 GHz, 128-bit, 4-stage Routers, 3 virt. networks w/ 3 virt. channels	1.5 GHz, 64x32x128bit Swizzle Switch Network
Main Memory	4096MB, 50 cycle latency	

- Benchmarks
 - SPLASH 2 : Scientific parallel application suite

Results—Performance & QoS

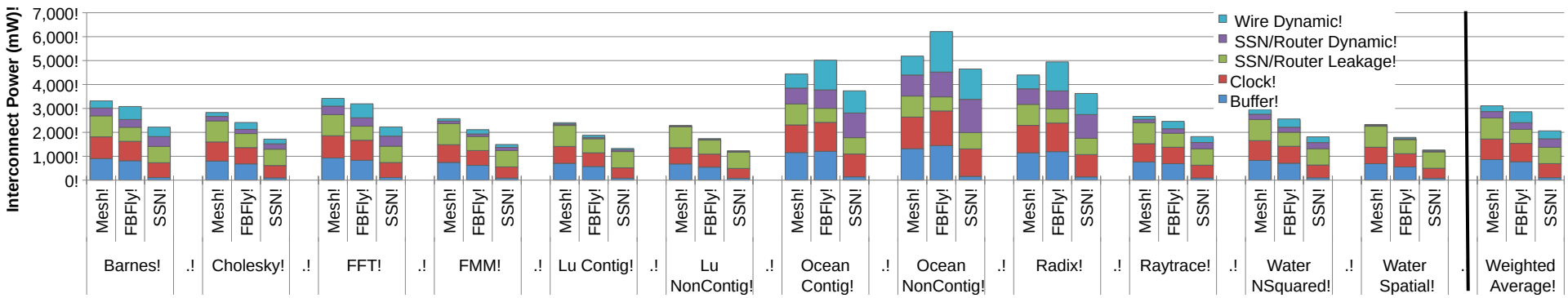


Overall Performance

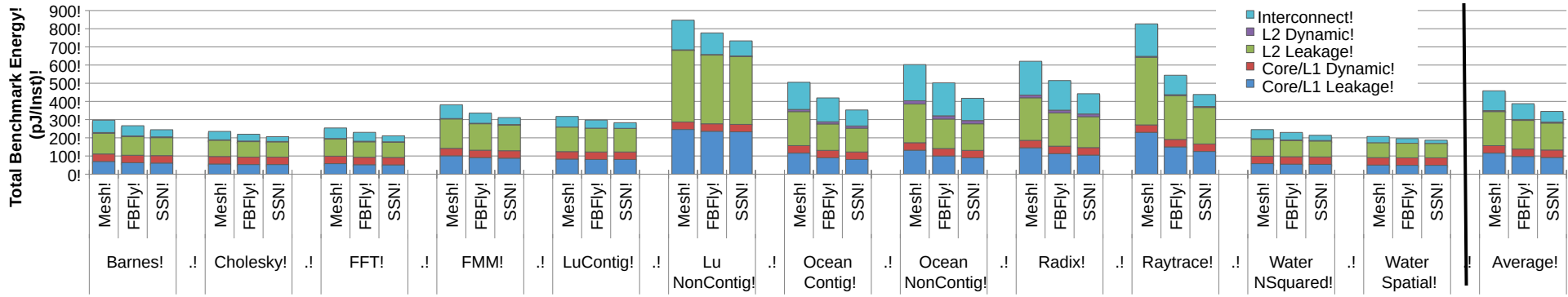


Quality-of-Service

Results—Power



On average the SSN uses **28%** less power in the interconnect compared to a flattened butterfly



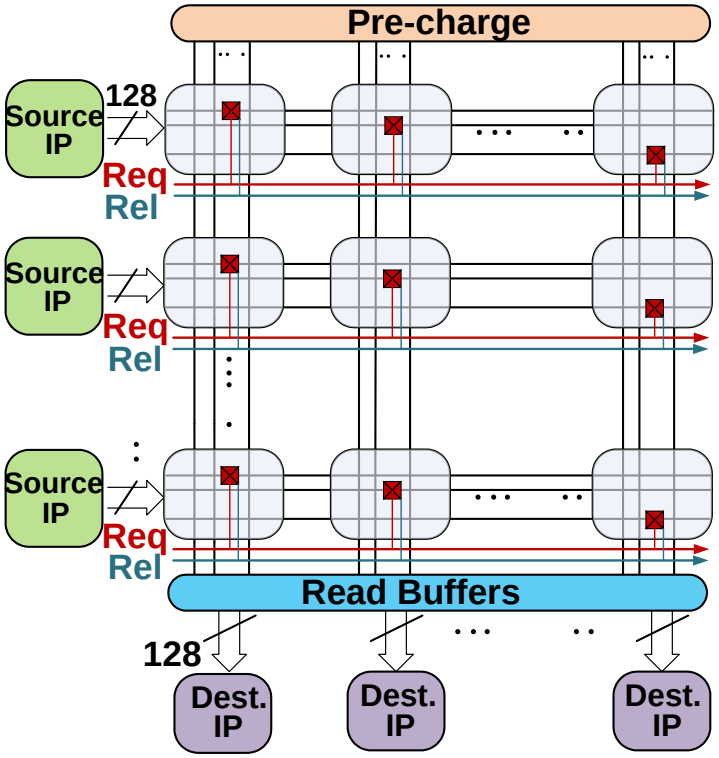
Which results in an average reduction in total system energy to complete the task of **11%**

Summary

- Swizzle Switch Prototype (45nm)
 - 64x64 Crossbar with 128-bit busses
 - Embedded LRG priority arbitration
 - Achieved 4.4 Tbps @ ~600MHz consuming only 1.3W of power
- Swizzle Switch Network Evaluation
 - Improved performance by 21%
 - Reduced power by 28%
 - Reduced latency variability by 3x

Additional Detailed Slides

Arbitration Mechanism (Matrix View)



Inhibits (X)

Requests (R)

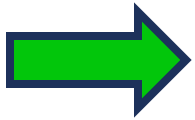
	X_0	X_1	X_2	X_3	X_4	*Priority*
R_0	X	1	0	0	0	***1***
R_1	0	X	0	0	0	***0***
R_2	1	1	X	0	0	***2***
R_3	1	1	1	X	1	***4***
R_4	1	1	1	0	X	***3***

Least Recently Granted (LRG)

	X_0	X_1	X_2	X_3	X_4	Priority
In_0	X	1	0	0	0	1
In_1	0	X	0	0	0	0
In_2	1	1	X	0	0	2
In_3	1	1	1	X	1	4
In_4	1	1	1	0	X	3

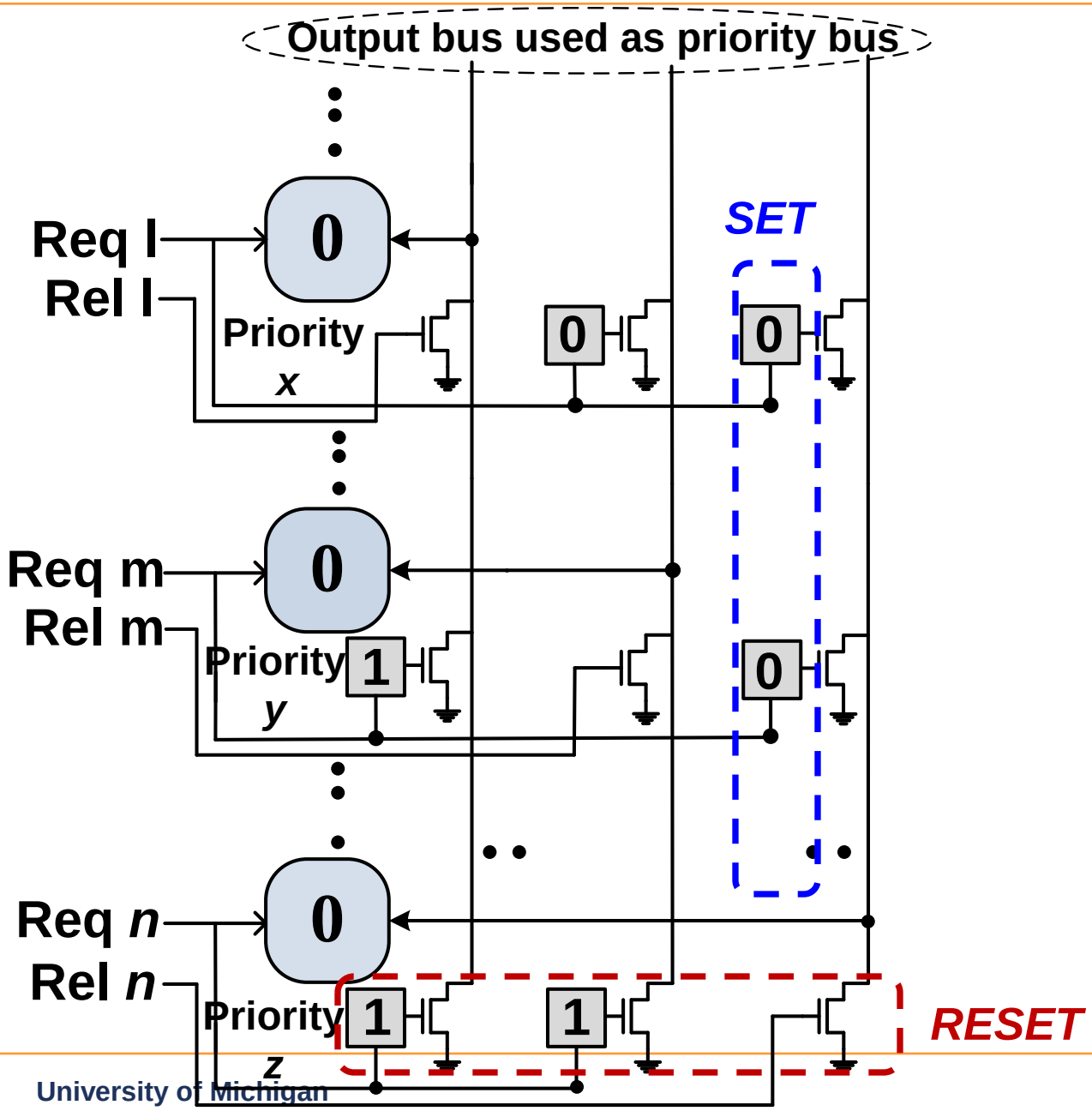
Set

Reset

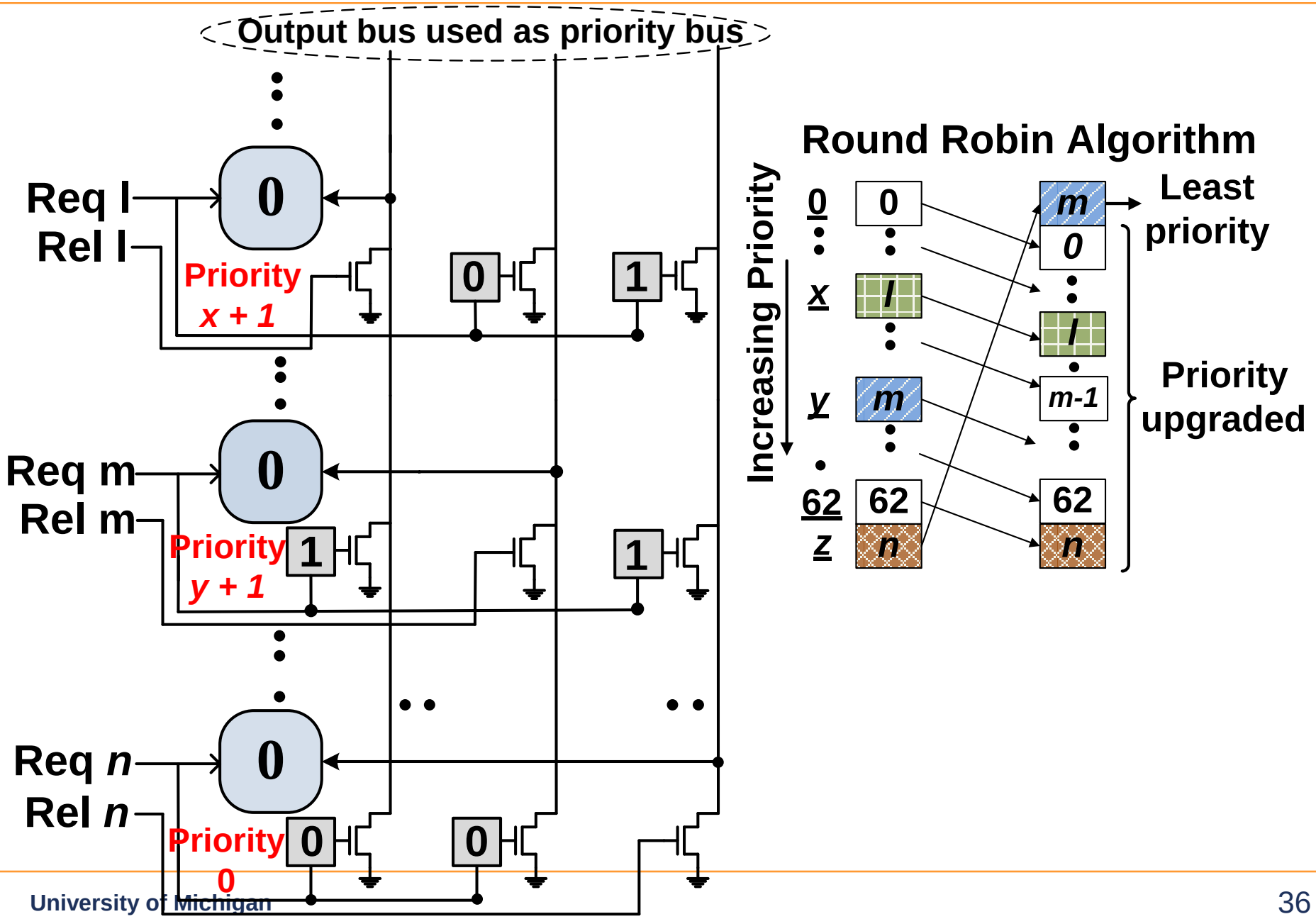


	X_0	X_1	X_2	X_3	X_4	Priority
In_0	X	1	0	0	1	2
In_1	0	X	0	0	1	1
In_2	1	1	X	0	1	3
In_3	1	1	1	X	1	4
In_4	0	0	0	0	X	0

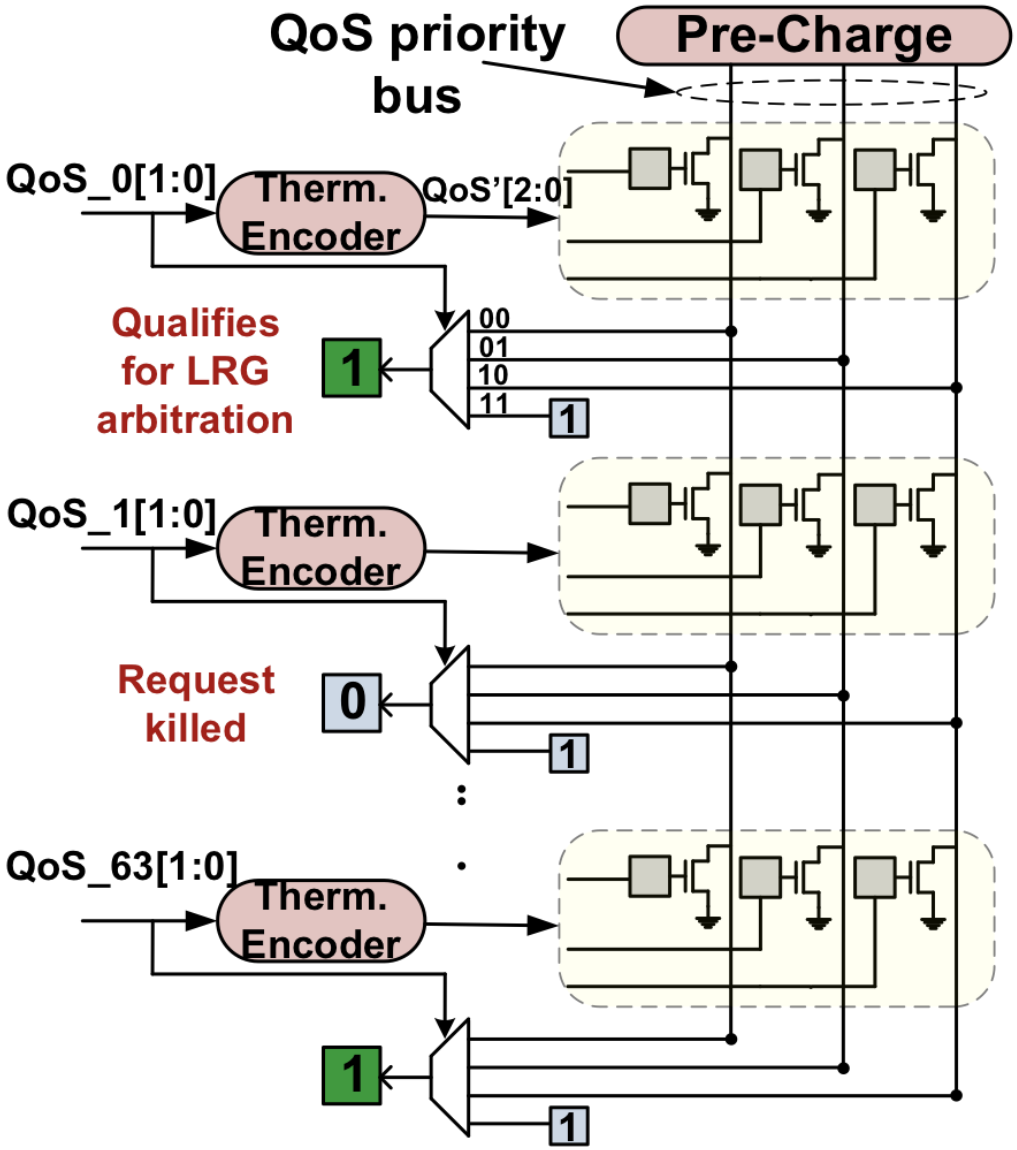
Round Robin Arbitration

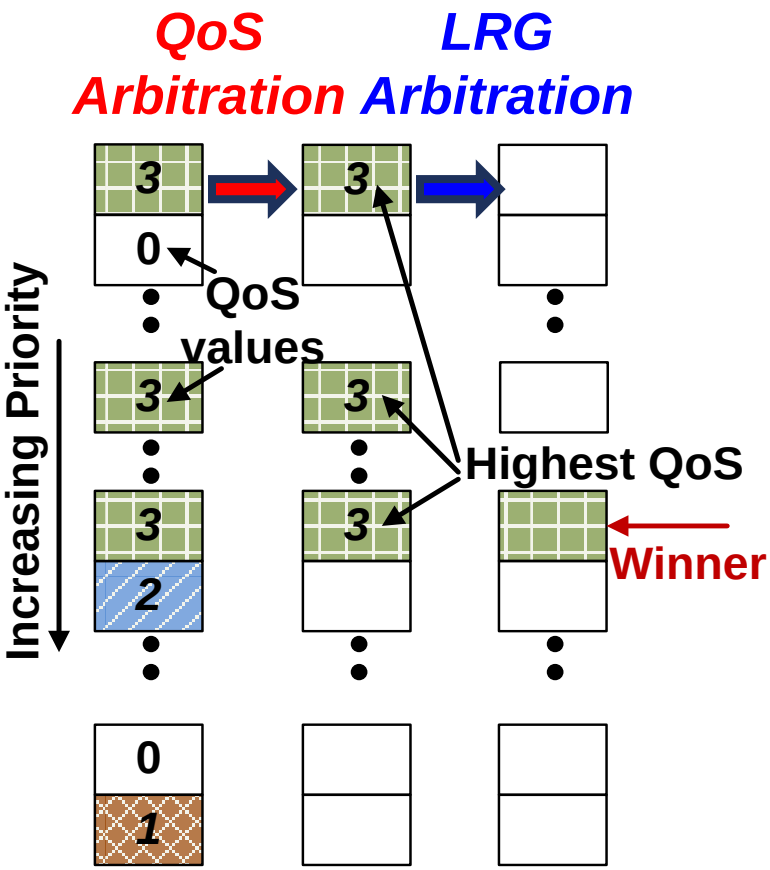


Round Robin Arbitration

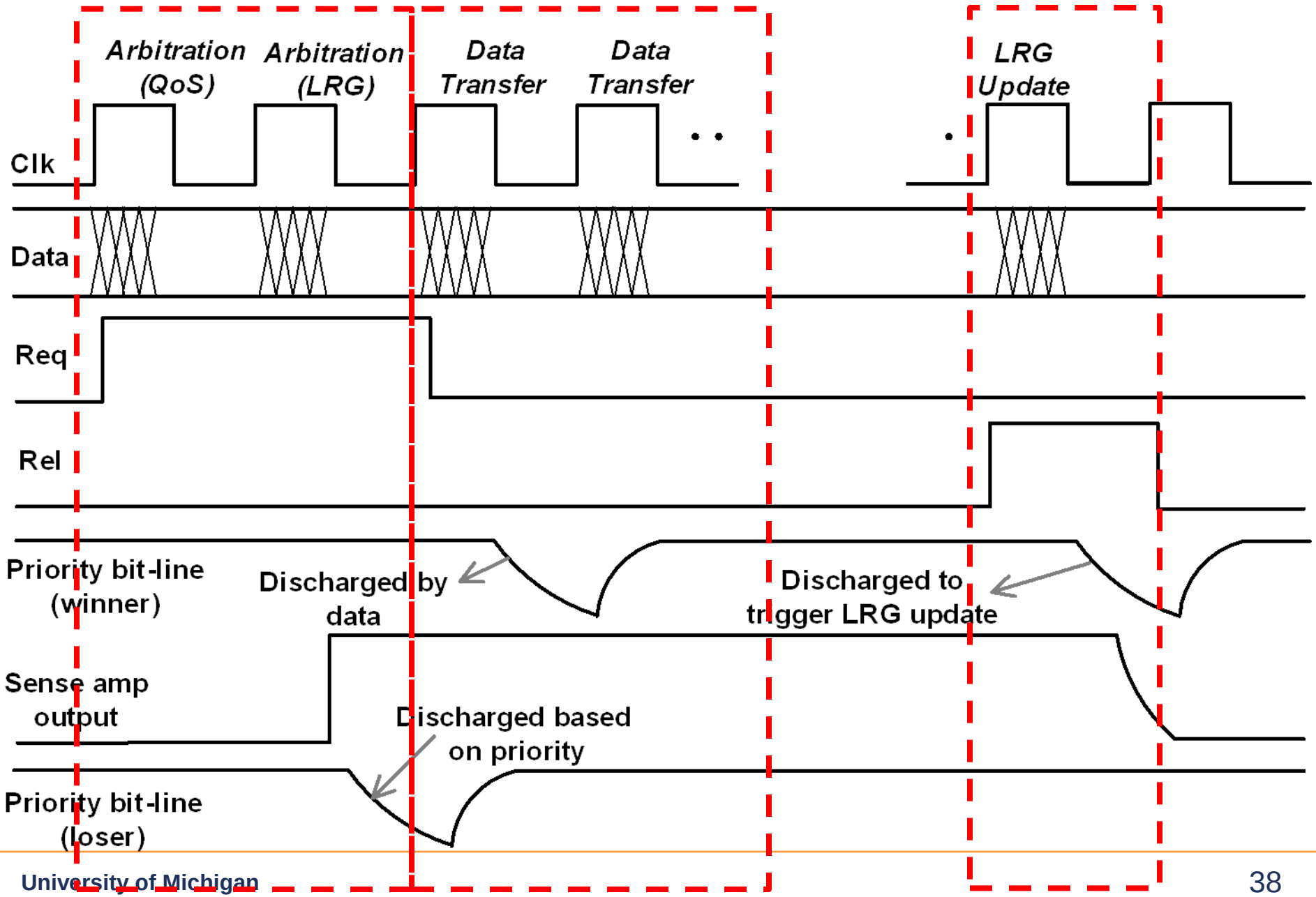


QoS Arbitration

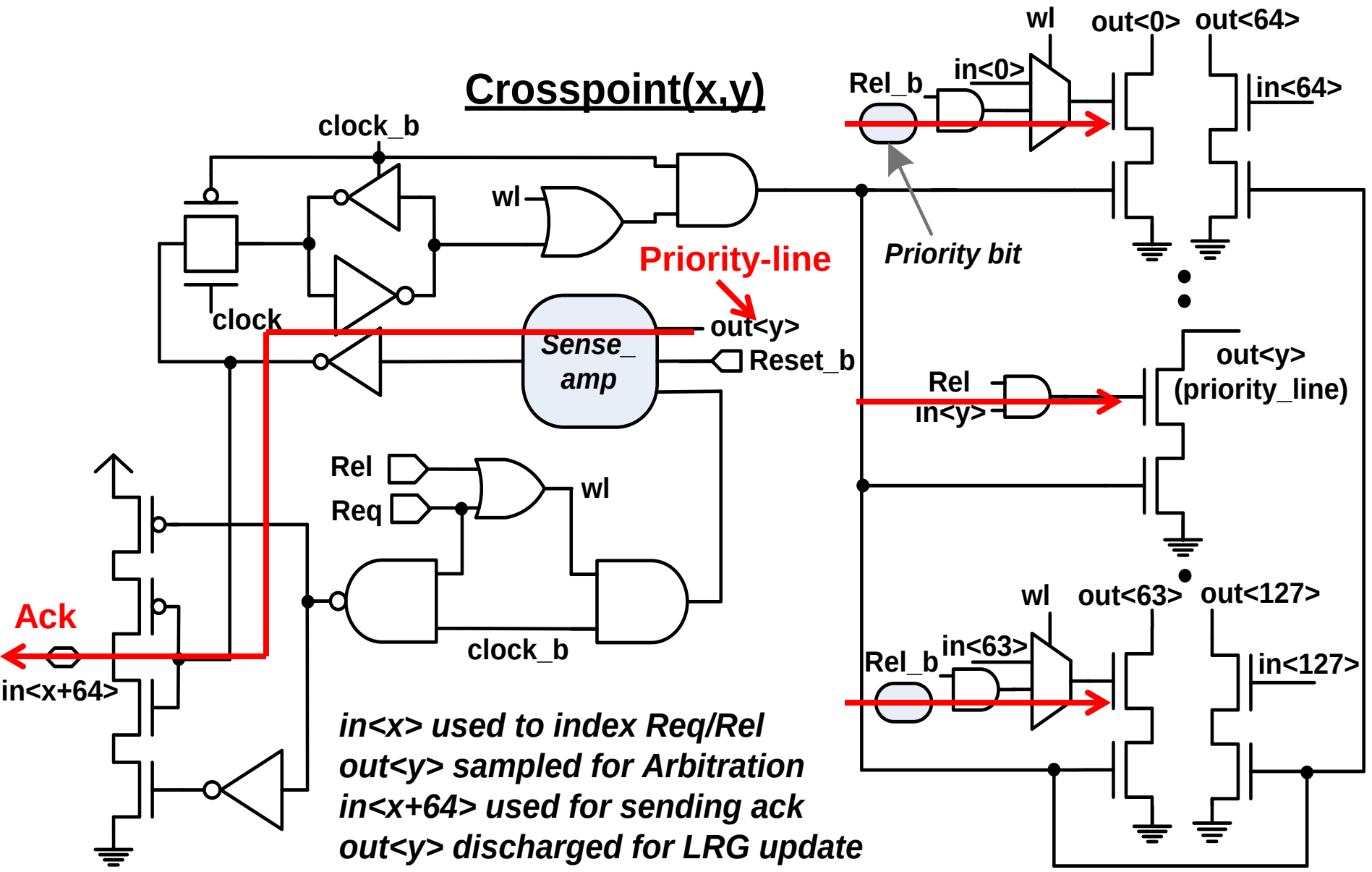




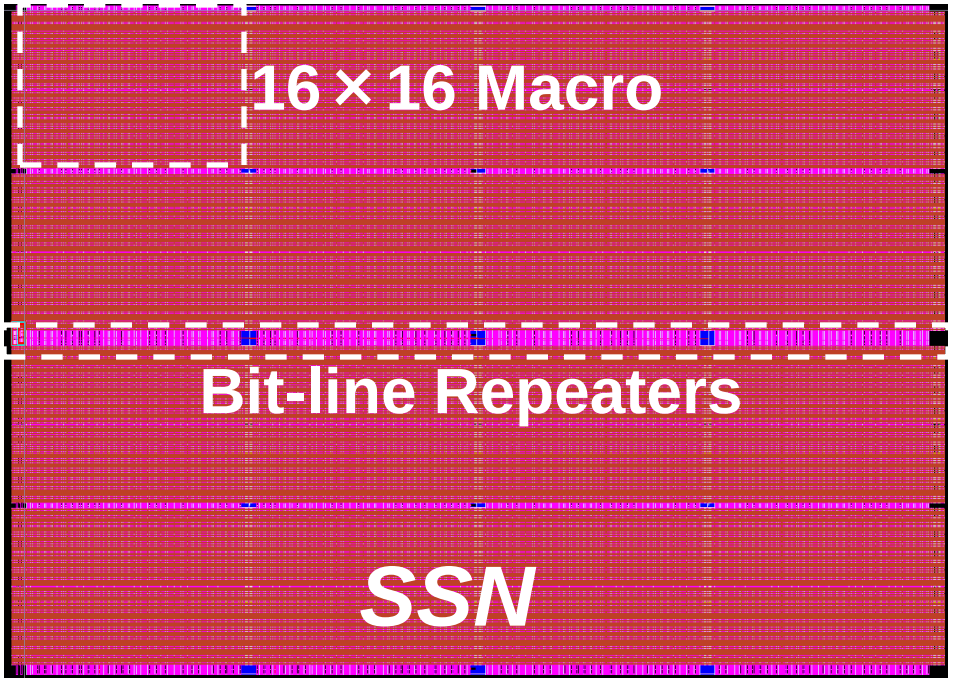
Timing Diagram

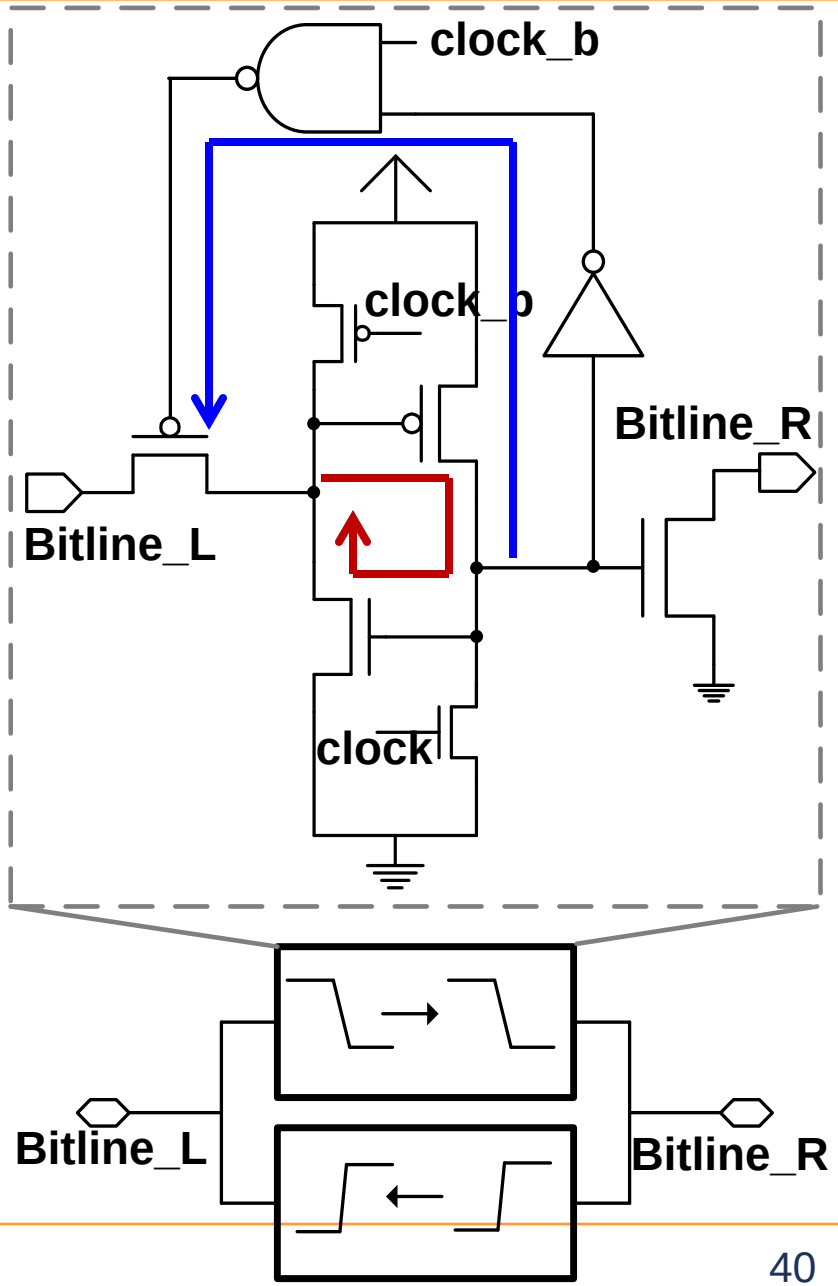


Crosspoint Circuit



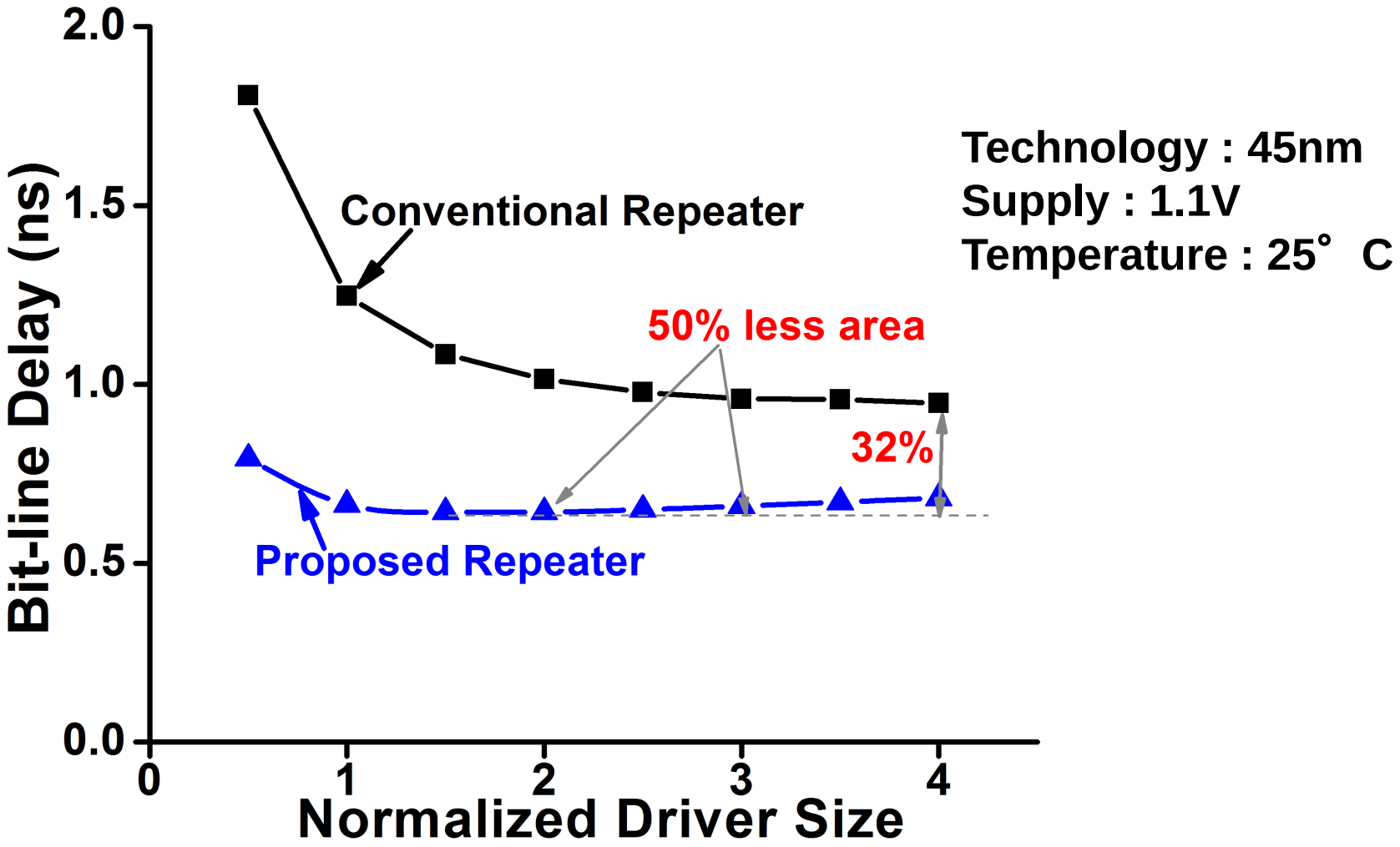
Regenerative Bit-line Repeater





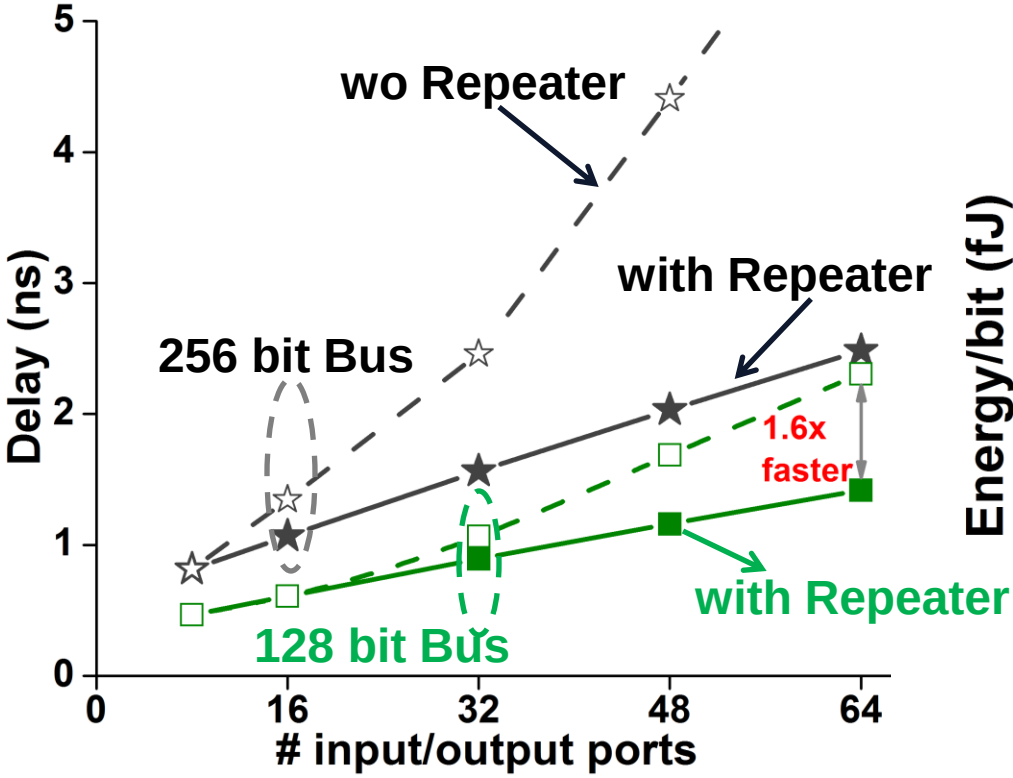
Regeneration and Decoupling improves speed

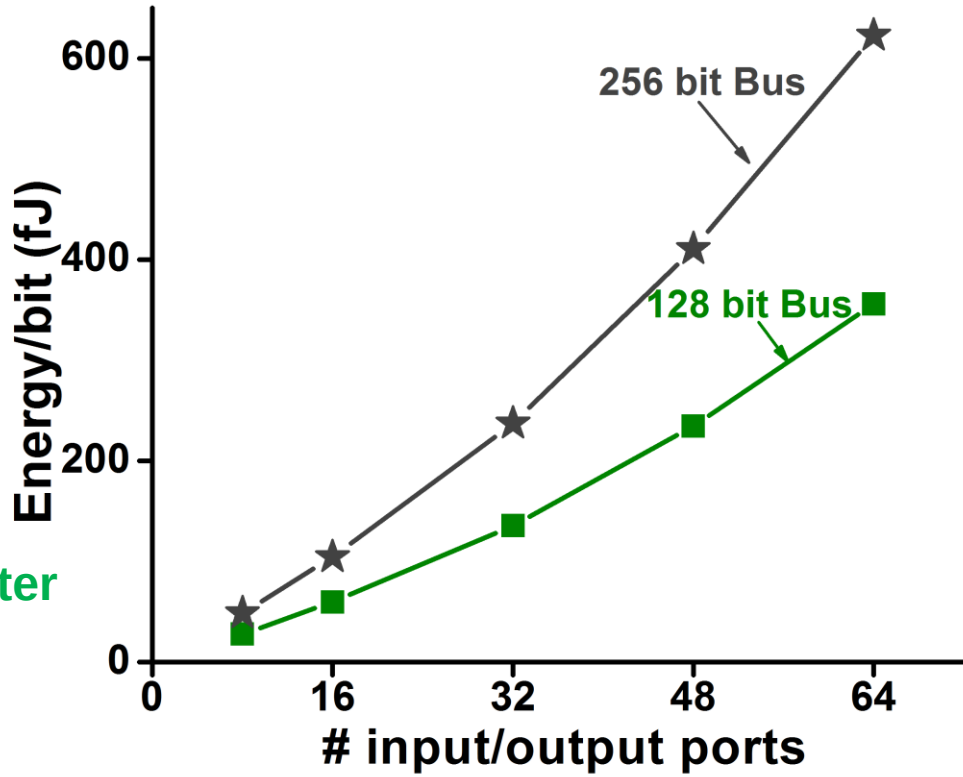
Simulated bit-line delay improvement



SSN Scaling: Simulation

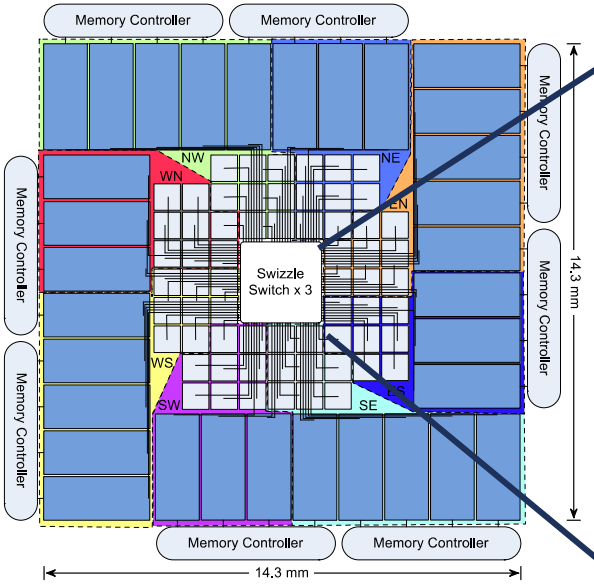
Technology : 45nm
Supply : 1.1V
Temperature : 25° C

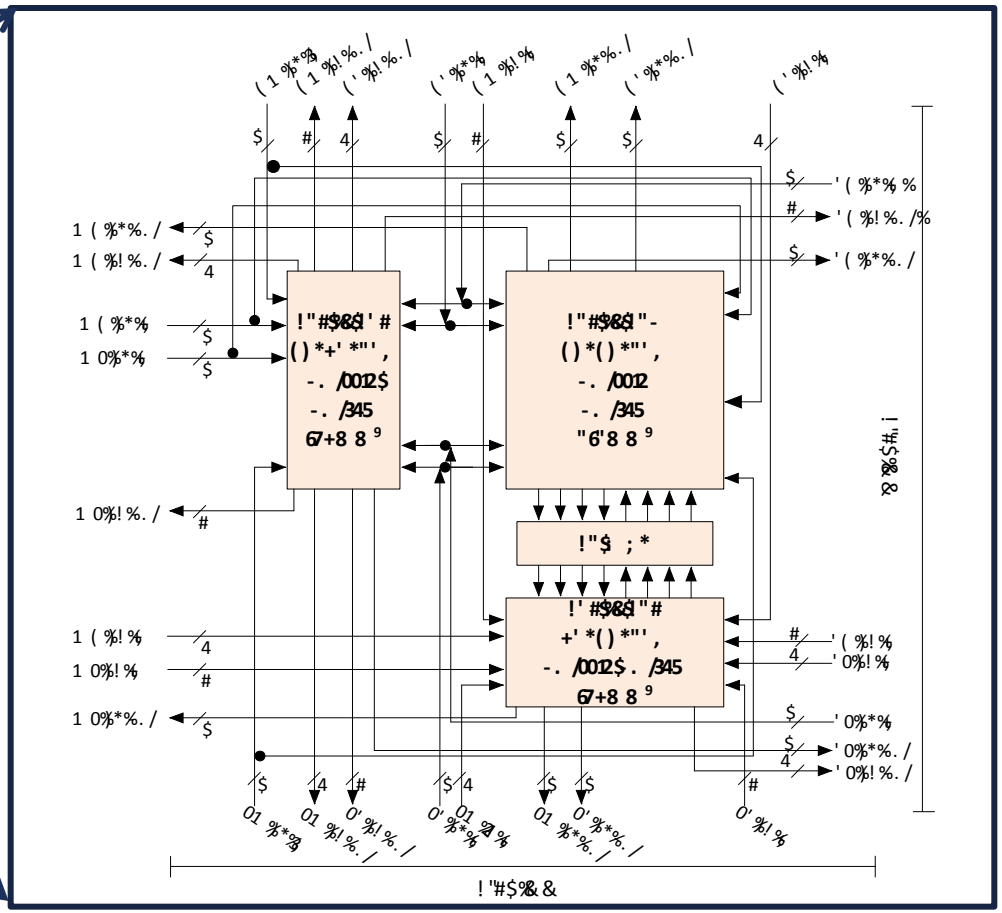




Regenerative repeaters improve SSN scalability

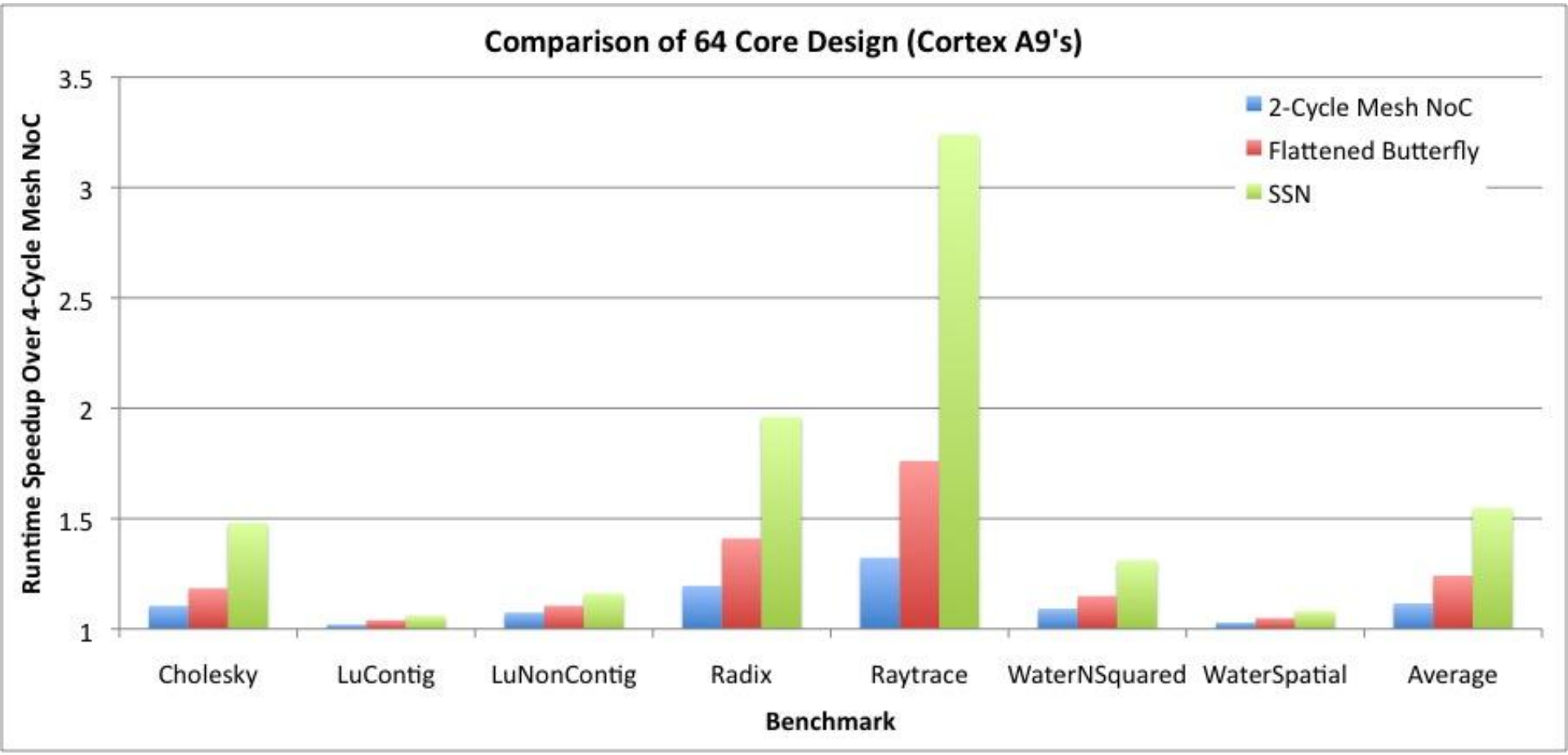
Swizzle Switch Network-on-Chip





Source \ Destination	L1	L2
L1	Shared Data Data Forwarding	Requests Writebacks
L2	Responses Invalidations	✕

Results—64-core with A9 O3 cores





FPGA Augmented ASICs: The Time Has Come

David Riddoch
Steve Pope



Copyright © 2012 Solarflare Communications, Inc. All Rights Reserved.

Hardware acceleration is Niche

- (With the obvious exception of graphics for gaming!)
- Even for people with a direct financial incentive to go faster…
- The reason?

It is enormously hard work!

- I'm going to talk about:
 - Why it is so hard
 - How to make it easier
 - Using online trading applications as an example
 - Industry perspective

Copyright © 2012 Solarflare Communications, Inc.

SOLARFLARE®

Trading applications

- Traders are in a latency race

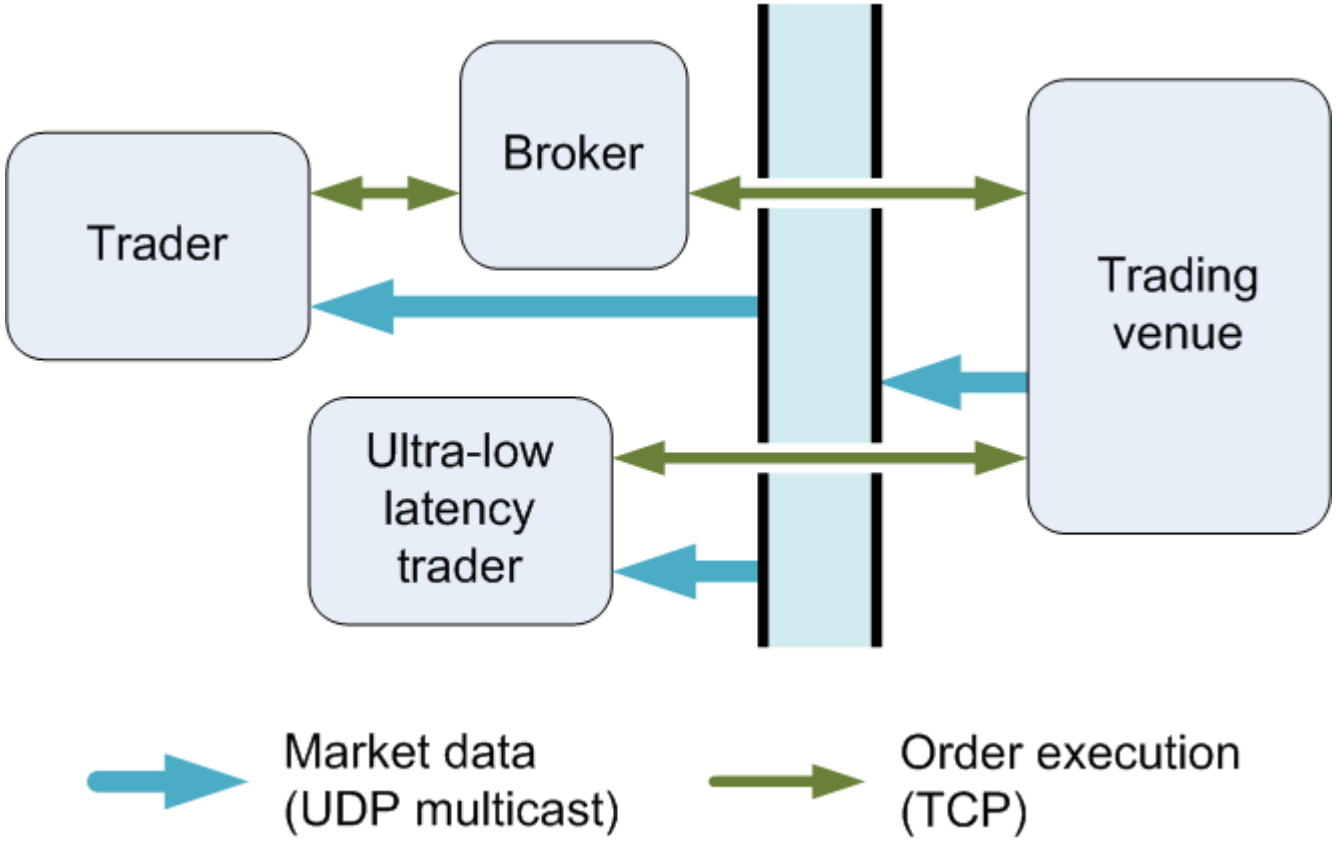
Signal ➔ Decision ➔ Action

- Whoever responds to signals fastest, makes money
- These folks have money to spend on technology, and exceptional engineers in-house
- But it is not a case of *performance-at-any-cost.* Like everyone else, they must balance performance against:
 - Flexibility / speed of deployment
 - Available skills
 - Cost
 - Compatibility

Copyright © 2012 Solarflare Communications, Inc.

SOLARFLARE®

Trading applications



…of course there is lots more that we don't have time to go into…

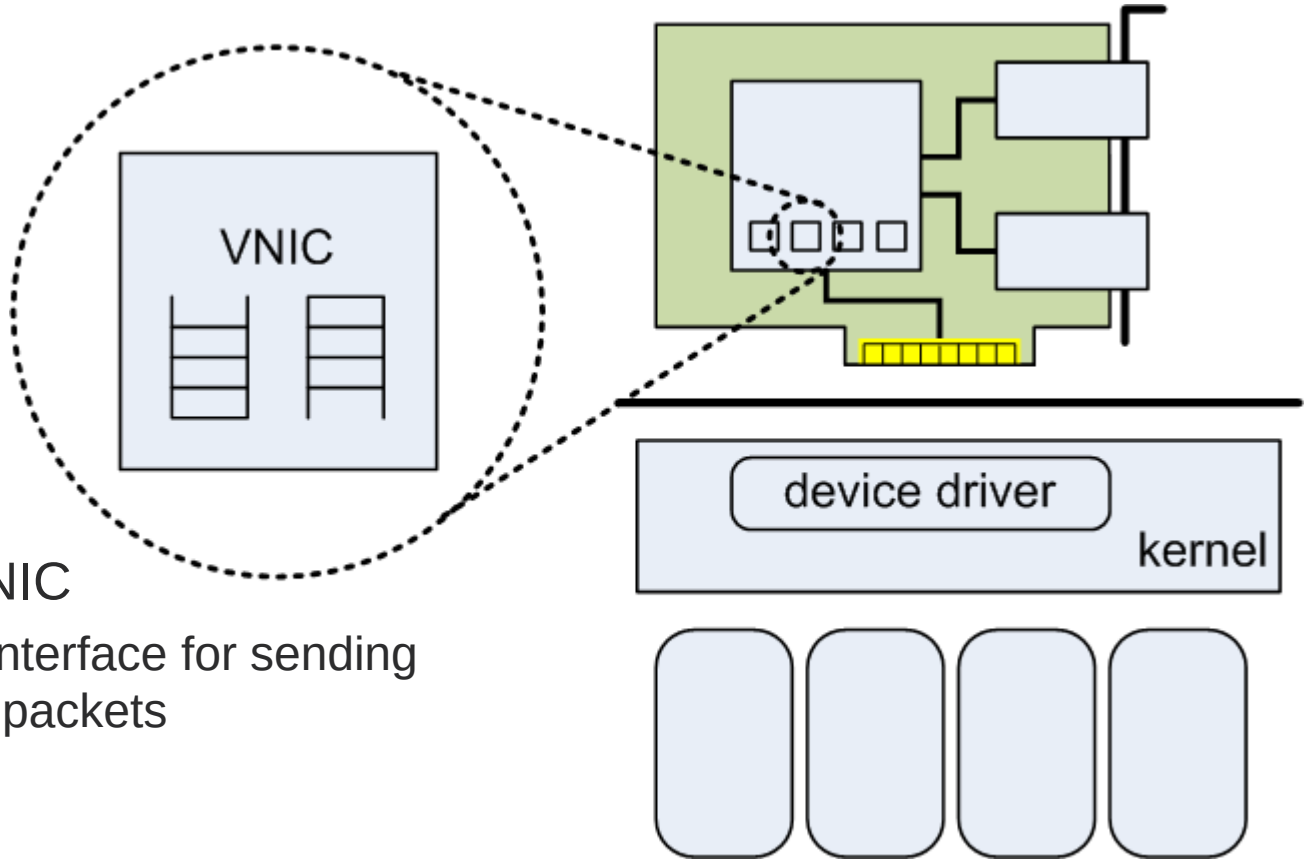
But all participants have a financial incentive to reduce latency, and many also have a throughput challenge.

Copyright © 2012 Solarflare Communications, Inc.

SOLARFLARE®

How does the NIC help?

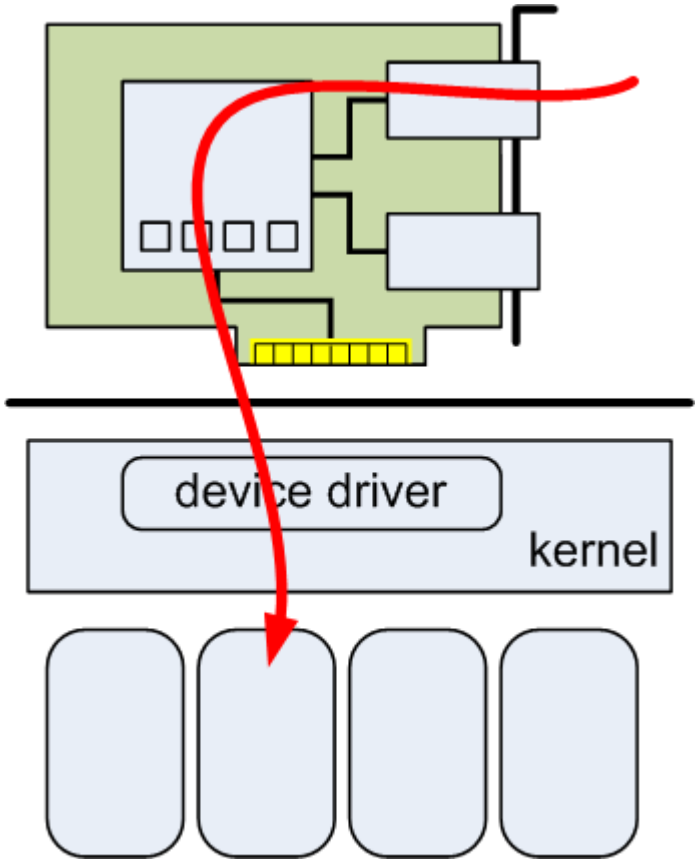
- Low latency cut-through design
- 1024 VNICs per port
- VNIC == Virtual NIC
 - Independent interface for sending and receiving packets
- Flow steering
 - Direct individual flows to specific VNICs
 - Supports scaling and NUMA locality



Copyright © 2012 Solarflare Communications, Inc. Slide 5

SOLARFLARE®

Kernel networking

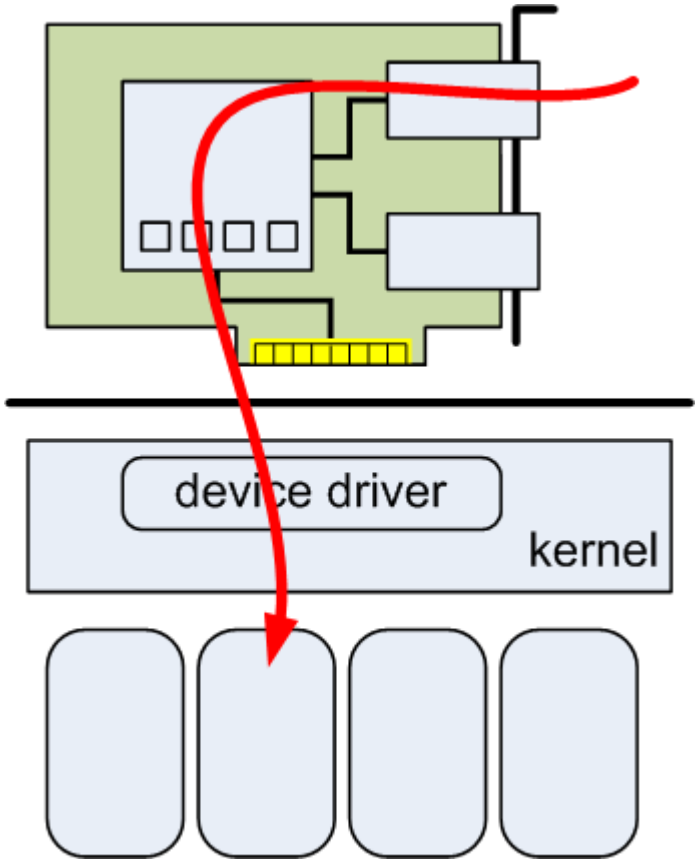


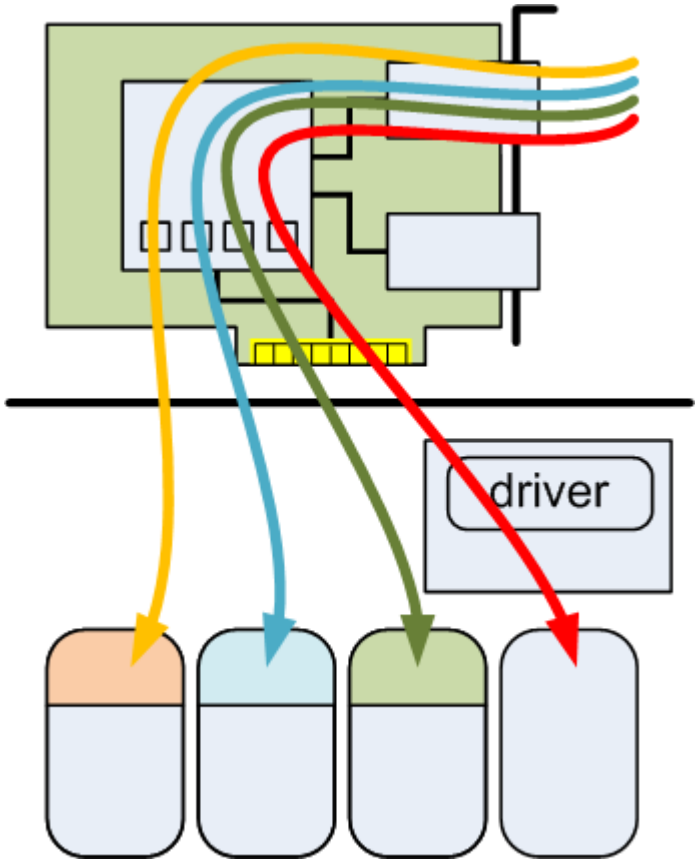
- Traditionally the network stack executes in the OS kernel
- Received packets are processed in response to interrupts
- Applications invoke the network via the BSD sockets interface by making system calls

Copyright © 2012 Solarflare Communications, Inc.

SOLARFLARE®

Kernel bypass



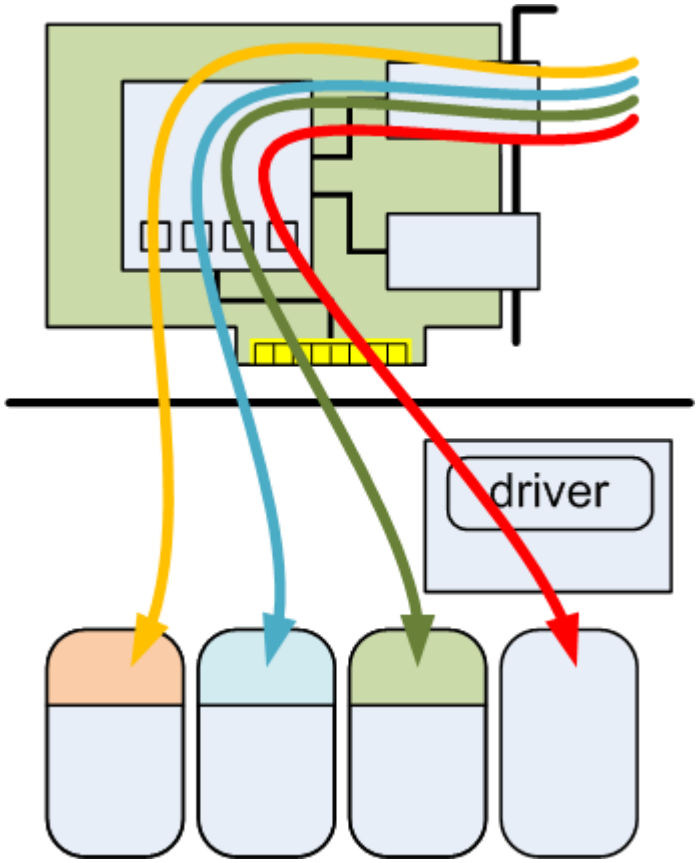


Copyright © 2012 Solarflare Communications, Inc. Slide 7

SOLARFLARE®

Kernel bypass – OpenOnload

- Dedicate a VNIC per application or thread
- TCP/UDP stack as user-level library
- Critical path entirely at user-level
- Reduces per-message CPU time
 - Cuts latency in half
 - Increases message rate by 5x per core
 - Improves scaling
- Fully compatible – no changes to applications needed



Copyright © 2012 Solarflare Communications, Inc.

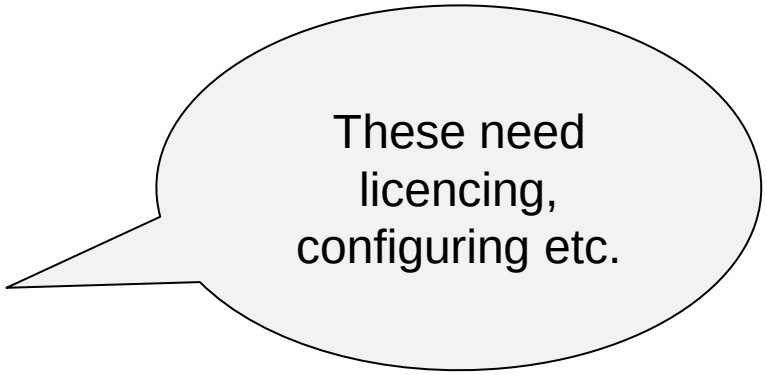
Back to the hardware…

Copyright © 2012 Solarflare Communications, Inc.

SOLARFLARE®

What is so hard about FPGA acceleration?

- Let's assume you want some custom logic
 - Evaluate the available board options
 - (Expensive ➔ low volume ➔ expensive ➔ low volume…)
- FPGA image:
 - Development tools
 - You'll need some IP blocks:
 - PCIe engine
 - Media access controller (MAC)
 - Memory controller
 - Boilerplate
 - Packet handling: Parsing, demultiplexing, buffering, streaming
 - Protocol handling: Checksums, headers, address resolution, TCP, UDP
 - Managing physical links: Configuration, errors, statistics, flow control
- Host software:
 - Device drivers
 - Control path
 - Fast interface to application (kernel bypass)



Copyright © 2012 Solarflare Communications, Inc.

SOLARFLARE®

What is so hard about FPGA acceleration?

- Let's assume you want some custom logic
 - Evaluate the available board options
 - (Expensive ➔ low volume ➔ expensive ➔ low volume…)
- FPGA image:
 - Development tools
 - You'll need some IP blocks:
 - PCIe engine
 - Media access controller (MAC)
 - Memory controller
 - Boilerplate
 - Packet handling: Parsing, demultiplexing, buffering, streaming
 - Protocol handling: Checksums, headers, address resolution, TCP, UDP
 - Managing physical links: Configuration, errors, statistics, flow control
- Host software:
 - Device drivers
 - Control path
 - Fast interface to application (kernel bypass)

We haven't written a single line of application logic yet!!!

These need licencing, configuring etc.

Copyright © 2012 Solarflare Communications, Inc.

SOLARFLARE®

So what is wrong with existing offerings?

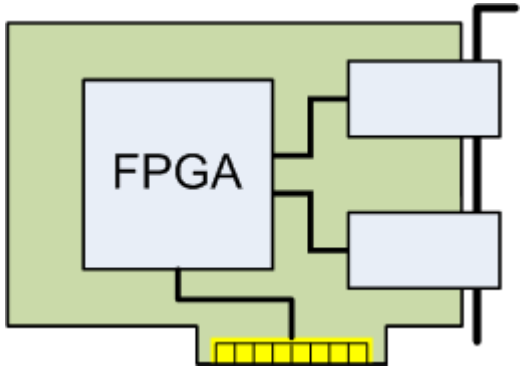
- Far too much work needed to create a deployable application
 - Apart from cost and time, the FPGA dev skills just aren't available
 - ➔ Need to provide the boilerplate (at the very least)
 - ➔ Need a much simpler host interface
- Hard to deploy incrementally
 - Requires simultaneous changes to multiple components
 - Accelerator network interface can only be used for accelerated traffic
 - Consumes an extra PCIe slot
 - ➔ Needs to integrate with existing apps
- Expensive
 - Requires huge benefit to justify investment
 - ➔ Need a solution that is widely useful (so we can make lots of them)
 - ➔ Need off-the-shelf applications to sell in volume

Copyright © 2012 Solarflare Communications, Inc.

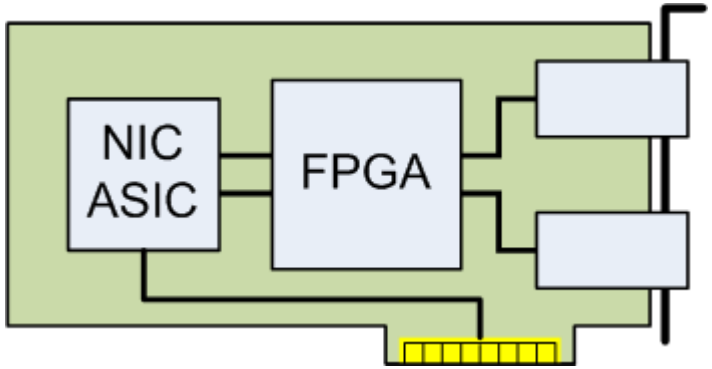
SOLARFLARE®

Solarflare's Application Onload Engine

- Not an FPGA with an Ethernet interface:



- A full-featured Ethernet adapter with FPGA accelerator:

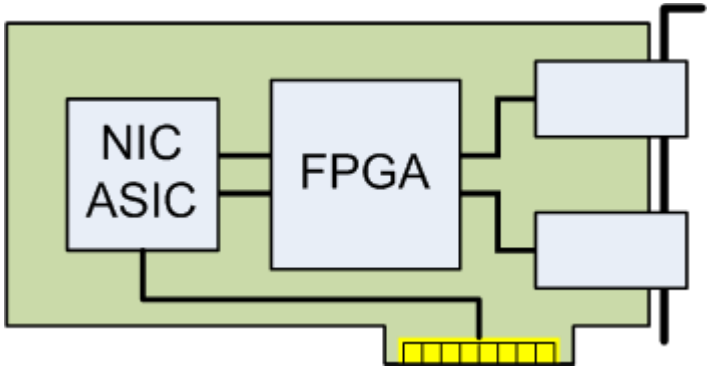


Copyright © 2012 Solarflare Communications, Inc.

Solarflare's Application Onload Engine

- Out of the box it works like a regular Solarflare network adapter
 - Drivers included
 - Works with kernel network stack and kernel bypass (OpenOnload)

- Incremental upgrade
 - Pass-thru by default
 - Accelerate a subset of traffic
 - No new switches, cabling, slot

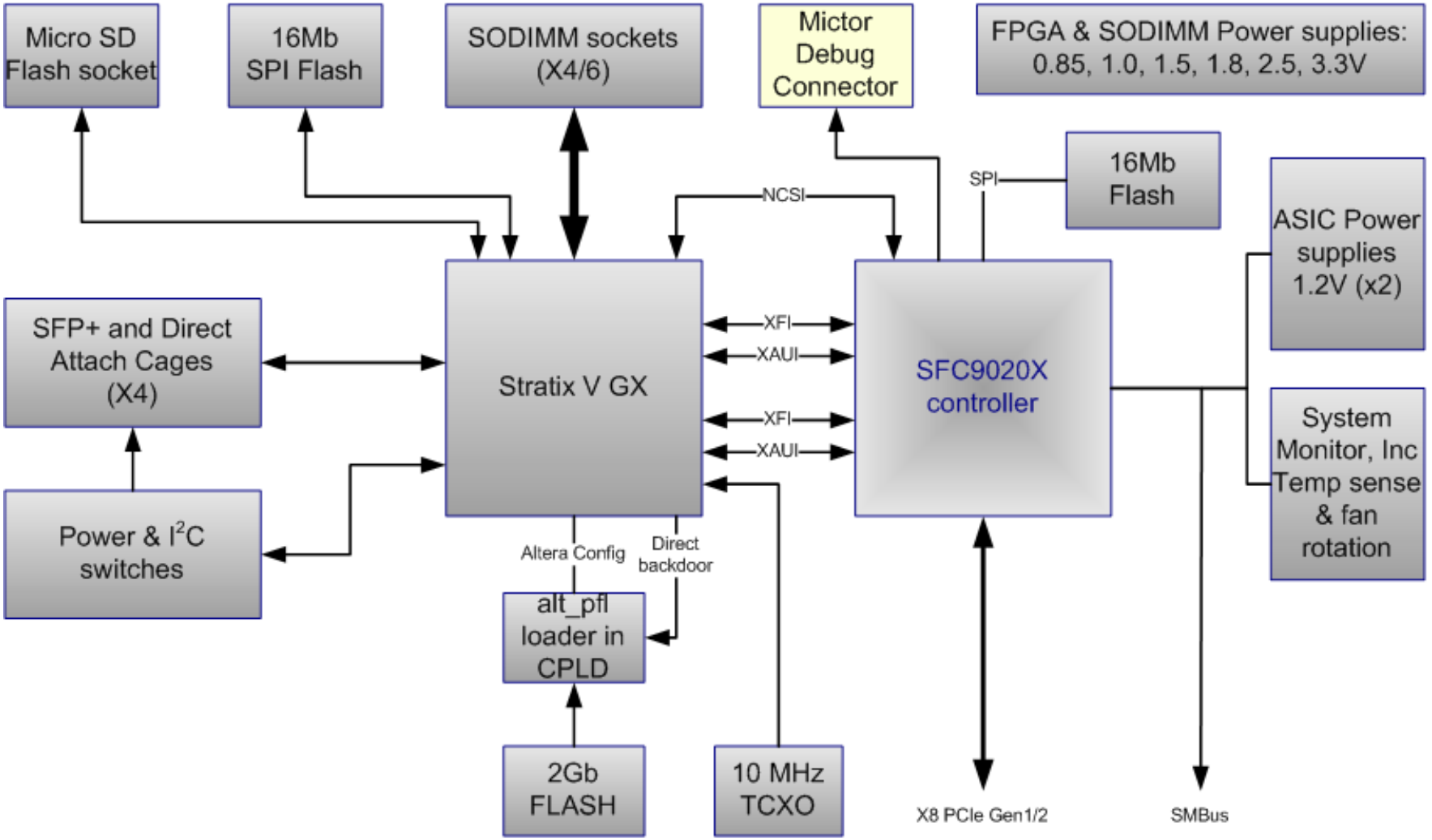


- Solarflare & 3rd party applications
 - Solve common problems
 - No FPGA expertise required

- FDK (developer kit)
 - Reusable IP blocks to minimise effort for FPGA developers

Copyright © 2012 Solarflare Communications, Inc.

AOE: Block diagram



Copyright © 2012 Solarflare Communications, Inc. Slide 15

Host software interface

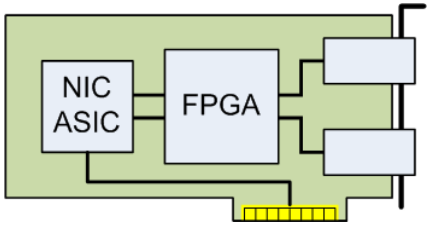
AOE™
ApplicationOnload™ Engine

- The data path is just *packets*
 - Applications on the host use BSD sockets
 - Via the kernel stack
 - Or via kernel bypass for higher performance
- We also provide a register bus
 - Mastered via a software API or command line tool
 - FPGA applications expose registers and memory
 - Notifications

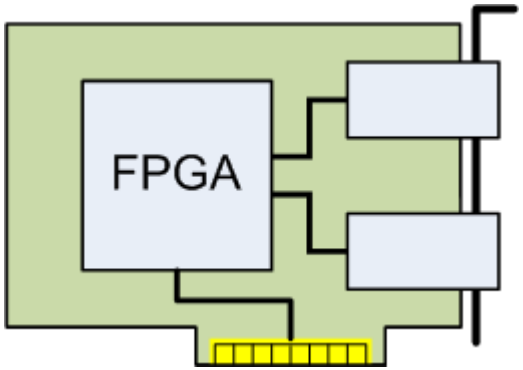
Copyright © 2012 Solarflare Communications, Inc.

SOLARFLARE®

What are the negatives?



- Compared to host-attached-FPGA boards:
 - AOE has higher latency between FPGA and host
 - AOE has no direct access to host from FPGA
 - Harder for FPGA apps to access host memory
 - Software on critical path
 - Much higher latency
 - FPGA can't master other devices

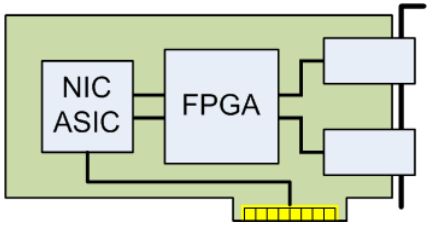


Copyright © 2012 Solarflare Communications, Inc. Slide 17

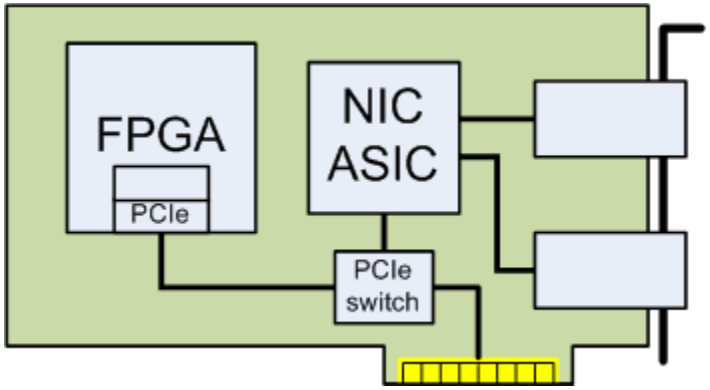
SOLARFLARE®

Alternative architectures?

- Bearing in mind we didn't want to change the ASIC



- PCIe attached FPGA
- Pros:
 - Fast access to host from FPGA
 - Better latency for pass-thru



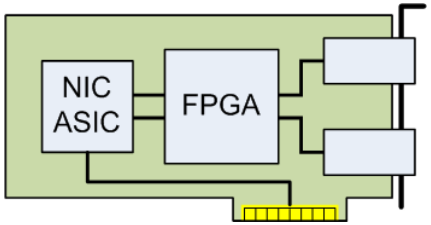
- Cons
 - Increased complexity
 - FPGA interacts with NIC ASIC via descriptor rings
 - Need new interface between FPGA and host
 - PCIe core in FPGA
 - (Less space for other things)

Copyright © 2012 Solarflare Communications, Inc.

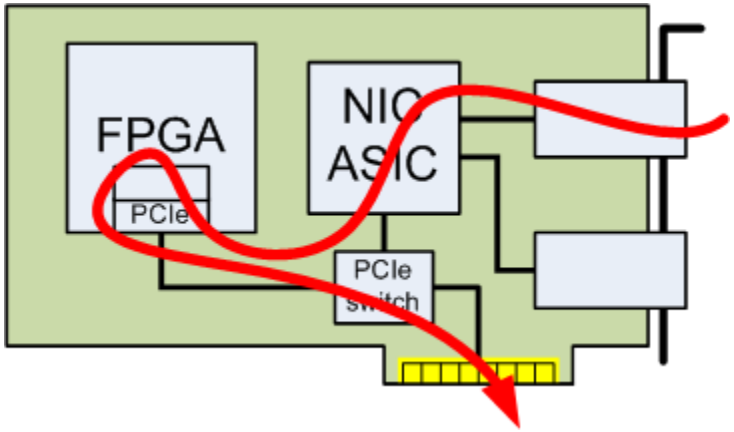
SOLARFLARE®

Alternative architectures?

- Bearing in mind we didn't want to change the ASIC



- PCIe attached FPGA
- Pros:
 - Fast access to host from FPGA
 - Better latency for pass-thru



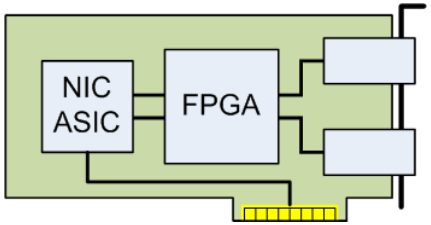
- Cons
 - Increased complexity
 - FPGA interacts with NIC ASIC via descriptor rings
 - Need new interface between FPGA and host
 - PCIe core in FPGA
 - (Less space for other things)
 - Significantly worse latency between host and wire

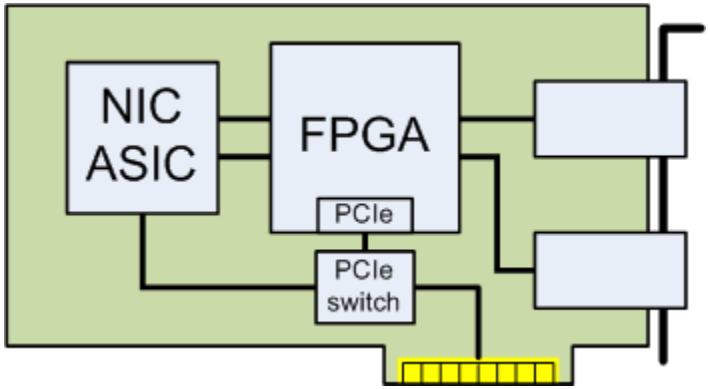
Copyright © 2012 Solarflare Communications, Inc.

SOLARFLARE®

Alternative architectures?

- Bearing in mind we didn't want to change the ASIC
- Add PCIe interface to FPGA
- Pros:
 - Fast access to host from FPGA
- Cons
 - Increased complexity
 - Need new interface between FPGA and host
 - PCIe core in FPGA
 - (Less space for other things)
 - Slightly worse latency through NIC



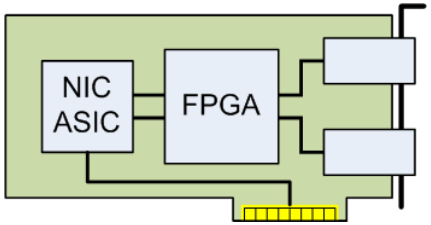


Copyright © 2012 Solarflare Communications, Inc. Slide 20

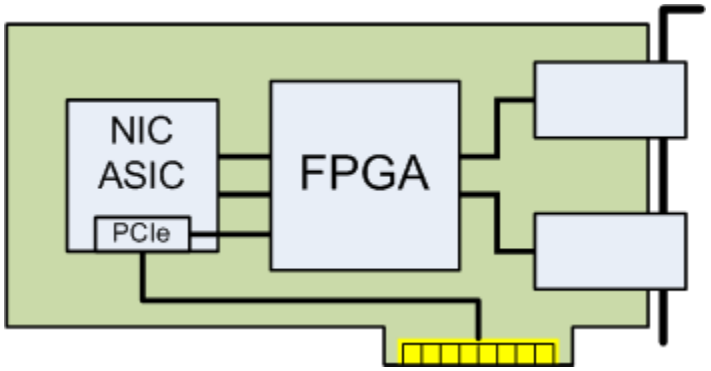
SOLARFLARE®

Alternative architectures?

- And if we could change the ASIC?



- Add fast bus for host access

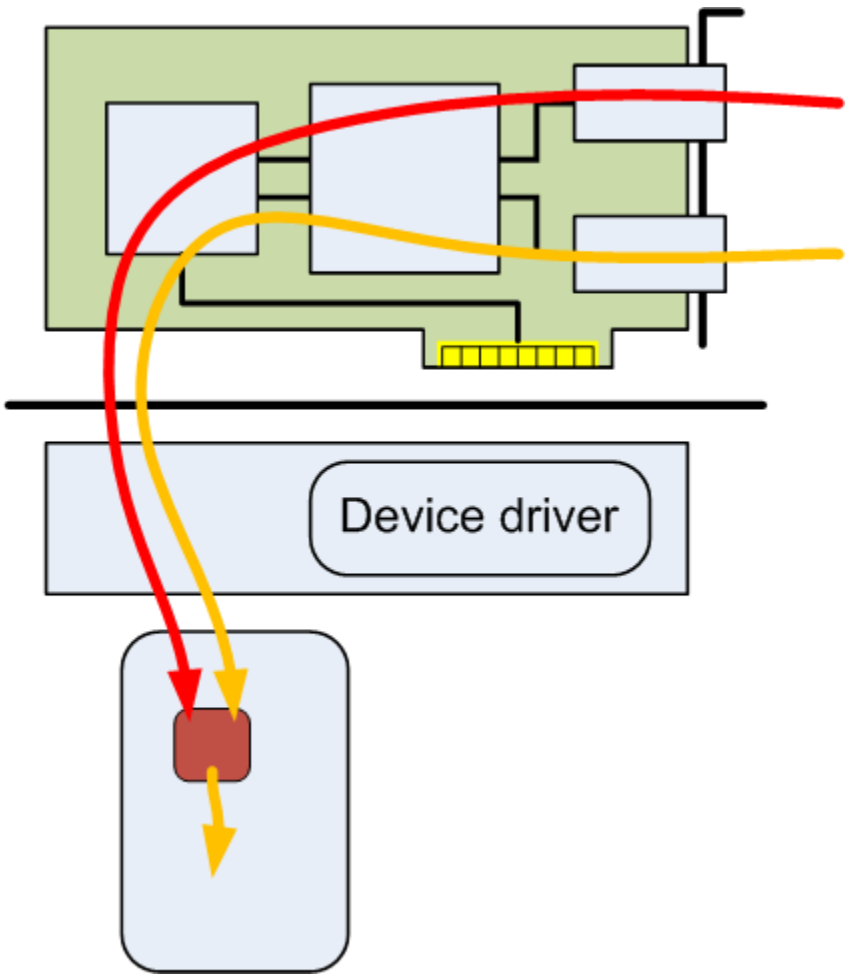


- Pros:
 - Fast access to host from FPGA
- Cons
 - Increased complexity
 - New interface between FPGA and host
 - But at least we're backwards compatible, so optional

Copyright © 2012 Solarflare Communications, Inc.

SOLARFLARE®

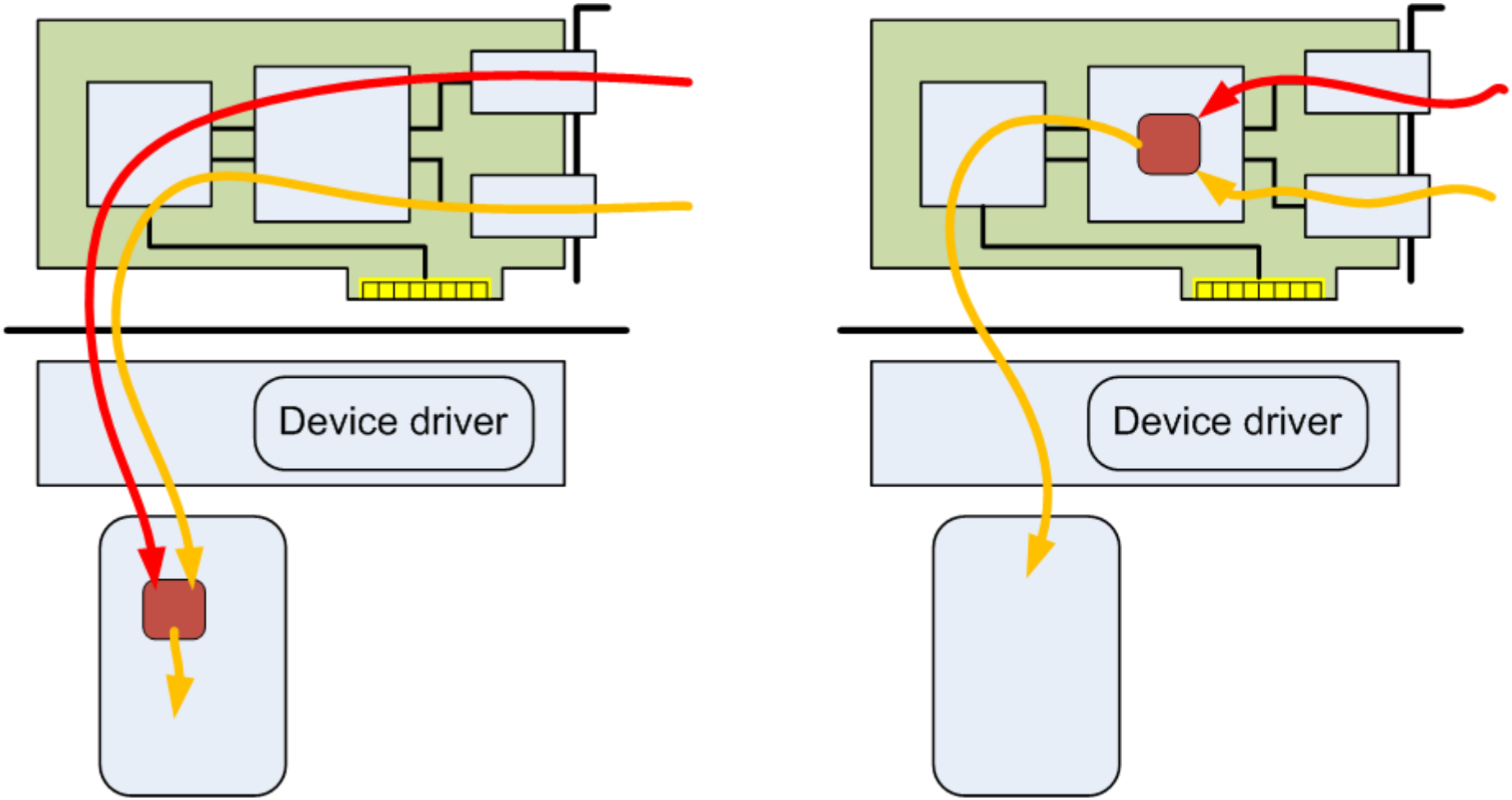
Example 1: Dual-line arbitration



- Market data is published as a pair of redundant feeds
- Traders often subscribe to both
 - For reliability
 - To get lowest latency
- Line arbitration converts the pair of streams into a single feed

Copyright © 2012 Solarflare Communications, Inc.

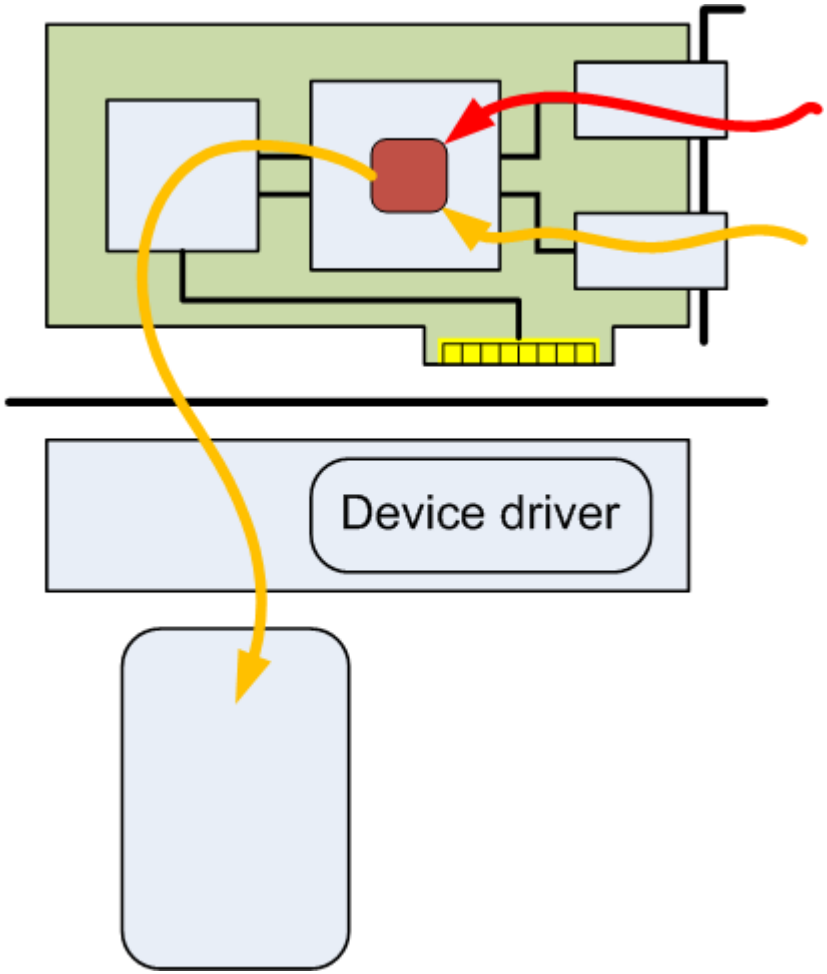
Example 1: Dual-line arbitration



Copyright © 2012 Solarflare Communications, Inc.

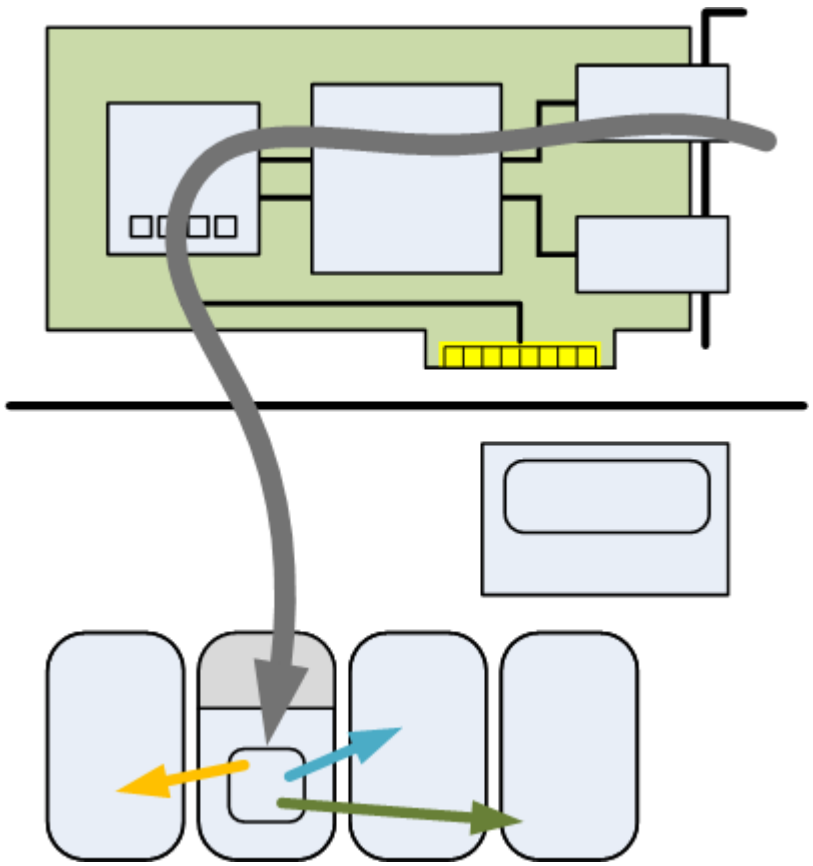
Example 1: Dual-line arbitration

- Line arbitration in the FPGA accelerator
 - Application sees a single stream
- Application gets the benefits of dual-line arbitration with half the data rate
- More likely to keep up
 - Reduces queuing delays
 - Reduces likelihood of unrecoverable loss due to buffer overflow
- No changes to software!



Copyright © 2012 Solarflare Communications, Inc.

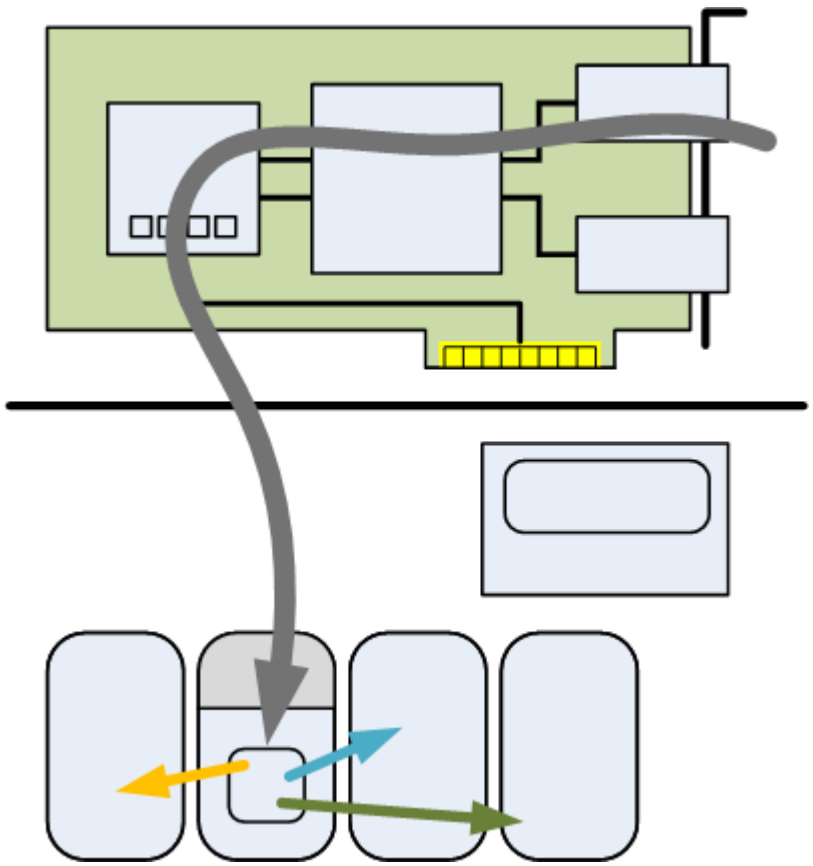
Example 2: Symbol splitting

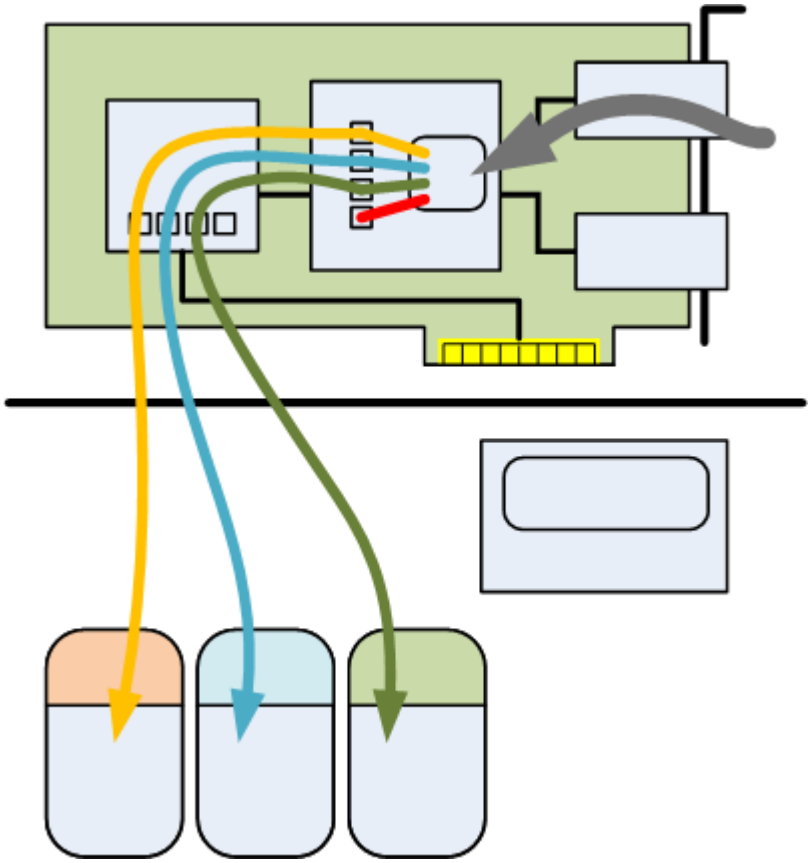


- Market data packets contain messages for multiple securities (symbols)
- Single or few packet streams
- Distributing load is a problem
 - May only be interested in a subset of symbols
 - Or may want to distribute load over multiple processes or threads
 - Must process messages in order (per symbol)
- Demultiplex in software is inefficient

Copyright © 2012 Solarflare Communications, Inc.

Example 2: Symbol splitting

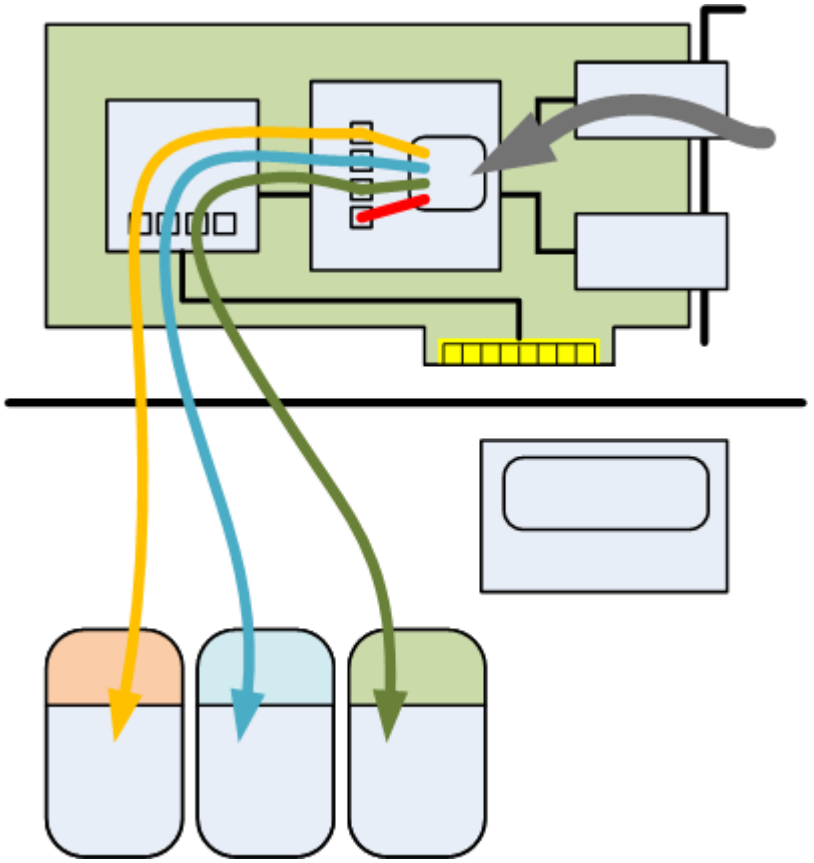




Copyright © 2012 Solarflare Communications, Inc. Slide 26

Example 2: Symbol splitting

- Split market data stream into per-symbol streams
- NIC distributes streams across processes, threads, cores
- Much higher throughput possible
 - More efficient because we've eliminated thread/cache interactions
- Lower latency
- Discard symbols we don't care about
 - Reduce throughput and queuing delays



Copyright © 2012 Solarflare Communications, Inc. Slide 27

Developer Kit Framework

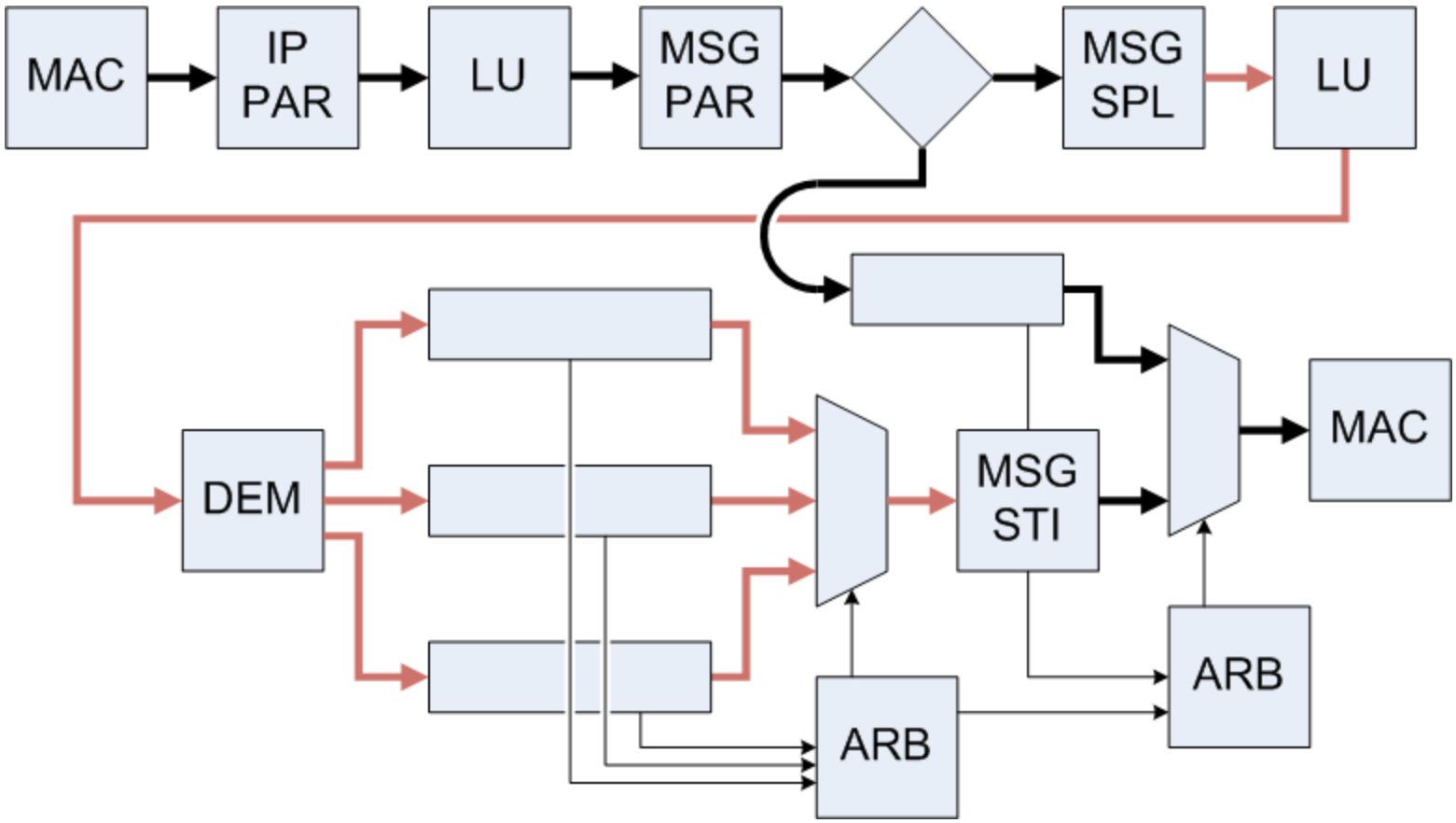
AOE™ ApplicationOnload™ Engine

- Next we show how the symbol splitter is implemented in the FPGA
- Using reusable components
- Connected by a streaming packet bus
 - Based on Altera's Avalon-ST streaming interface
 - Carries packets and/or messages
 - Meta-data words are interleaved within packets
- Components connected by packet bus may:
 - Inspect packets and add meta-data
 - Mutate packet data and meta-data
 - Pass-thru meta-data they don't recognise
 - Take actions based on meta-data
 - Manipulate state (lookup-tables, databases)
 - Routing decisions
 - Buffering (FIFOs, off-chip memory)

Copyright © 2012 Solarflare Communications, Inc.

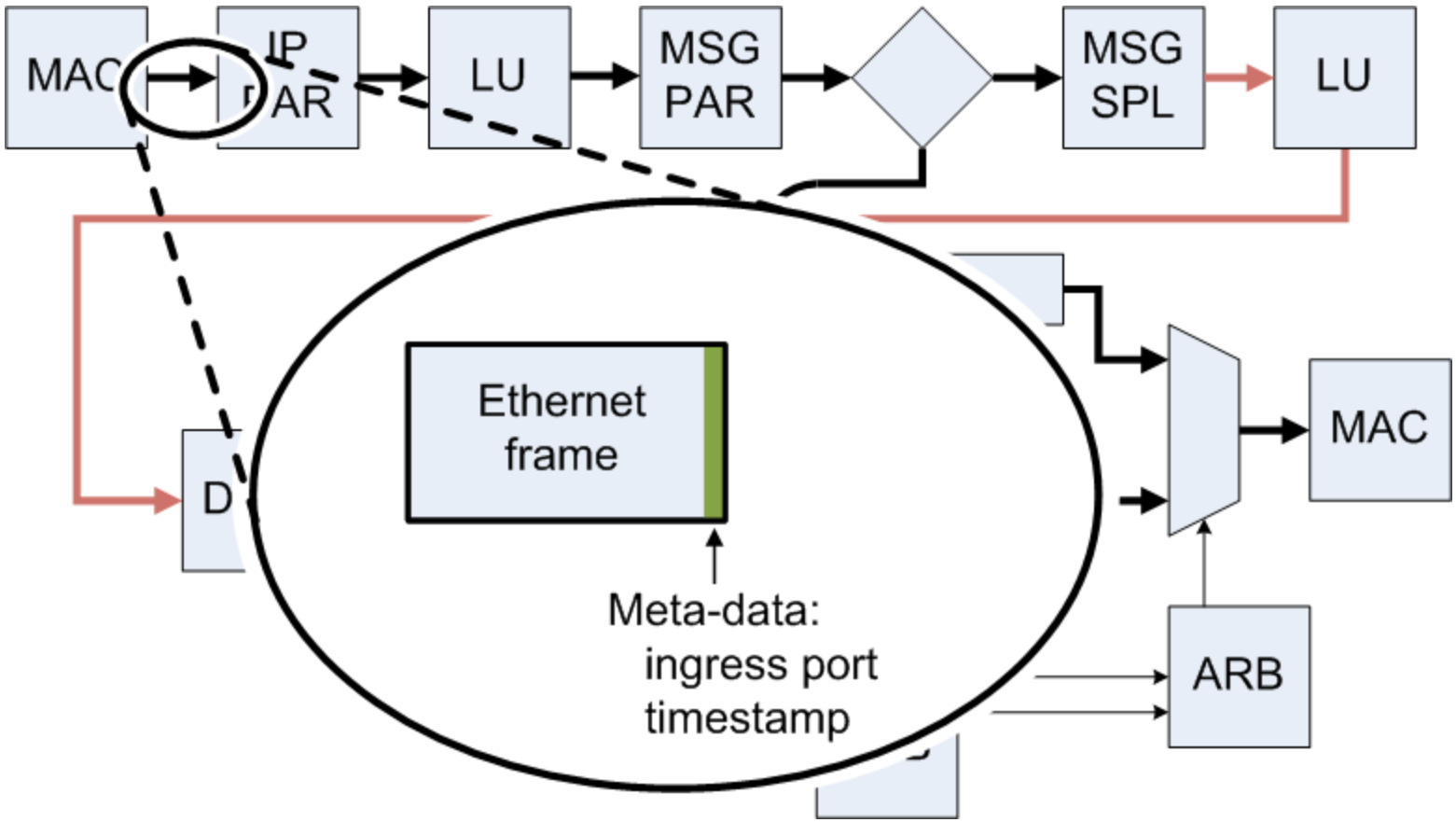
SOLARFLARE®

Symbol splitter implementation



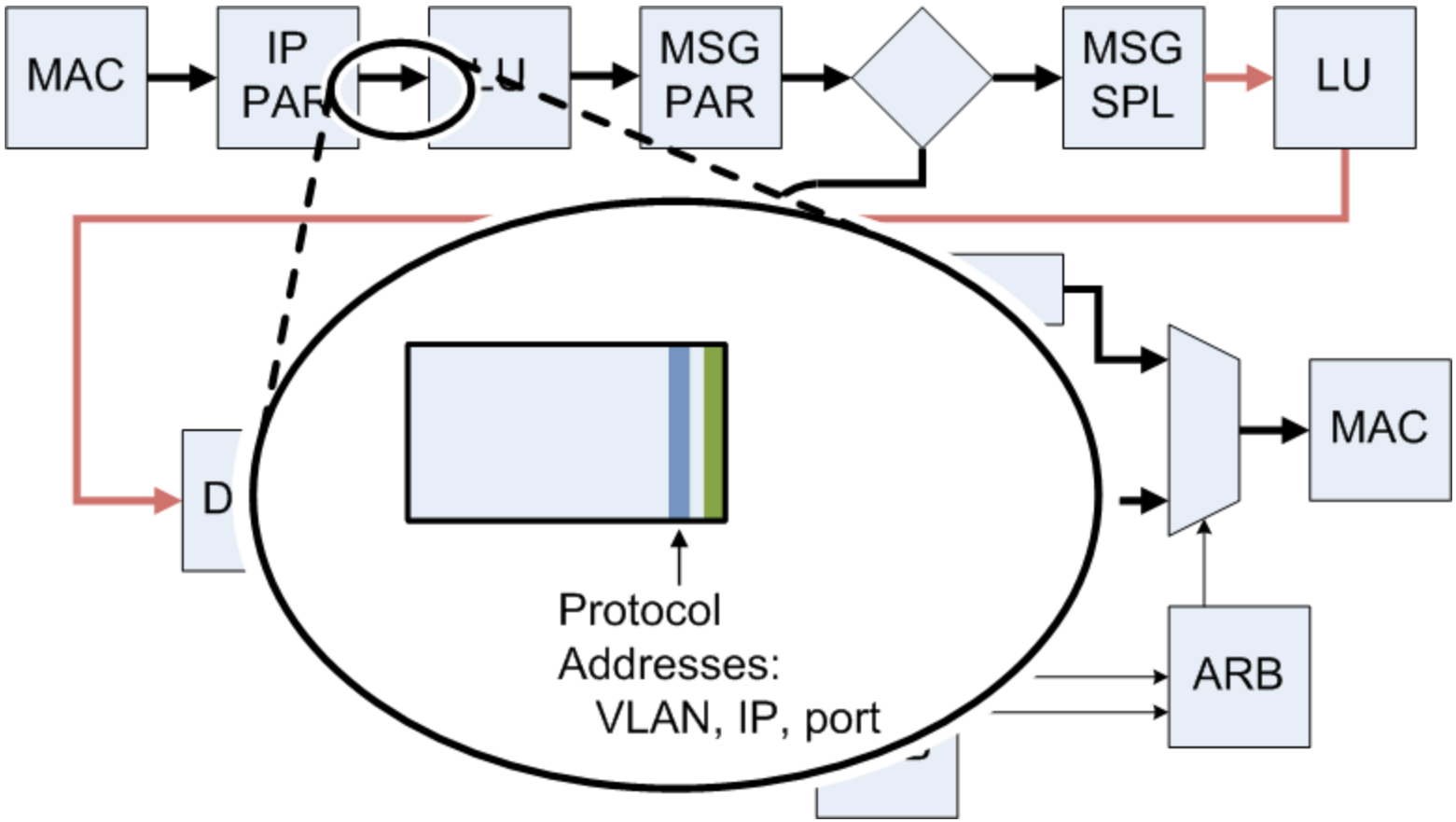
Copyright © 2012 Solarflare Communications, Inc. Slide 29

Single meta-data word a start of each packet



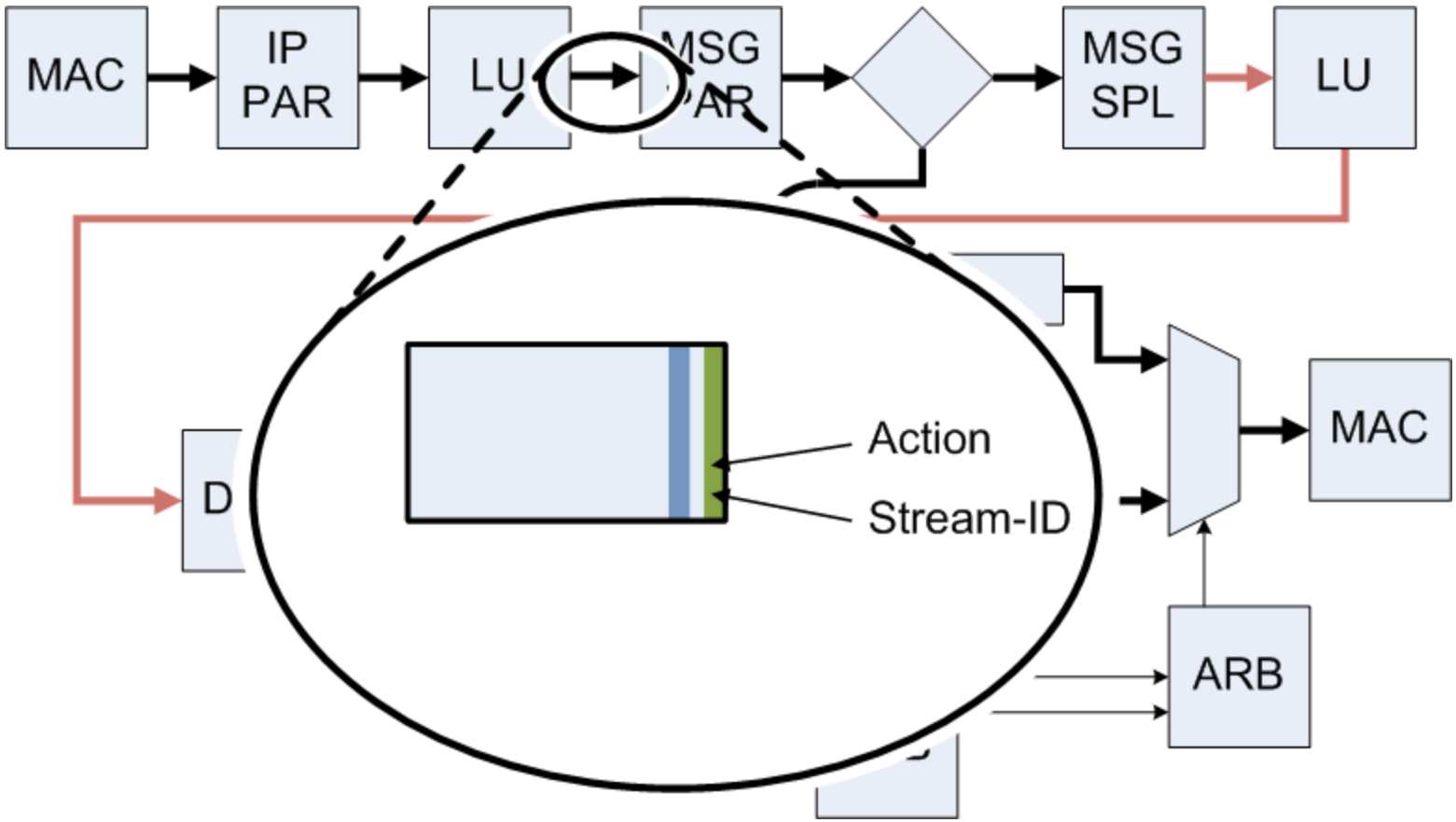
Copyright © 2012 Solarflare Communications, Inc. Slide 30

Parse headers, add meta-data



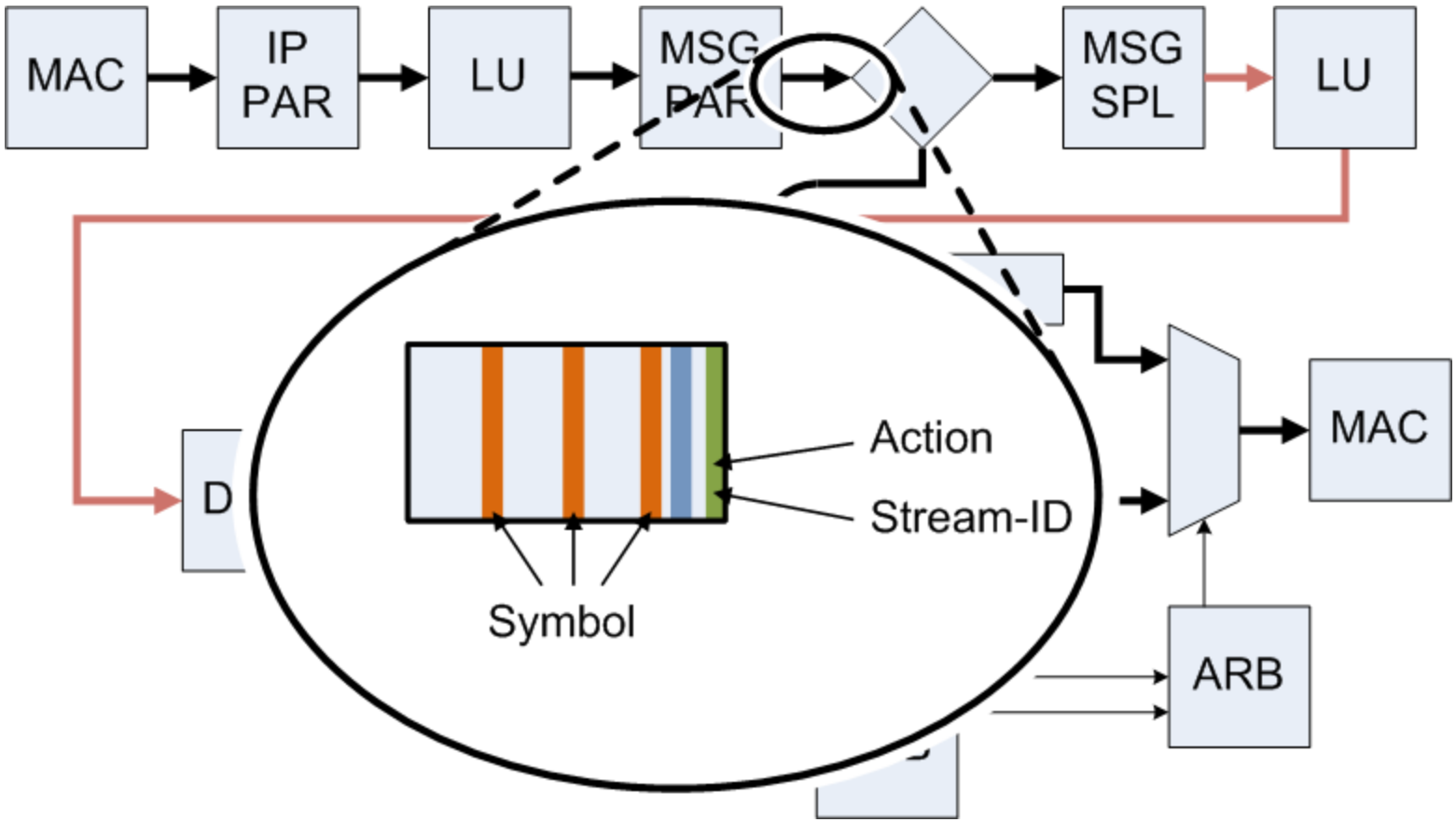
Copyright © 2012 Solarflare Communications, Inc. Slide 31

Lookup steam, add meta-data



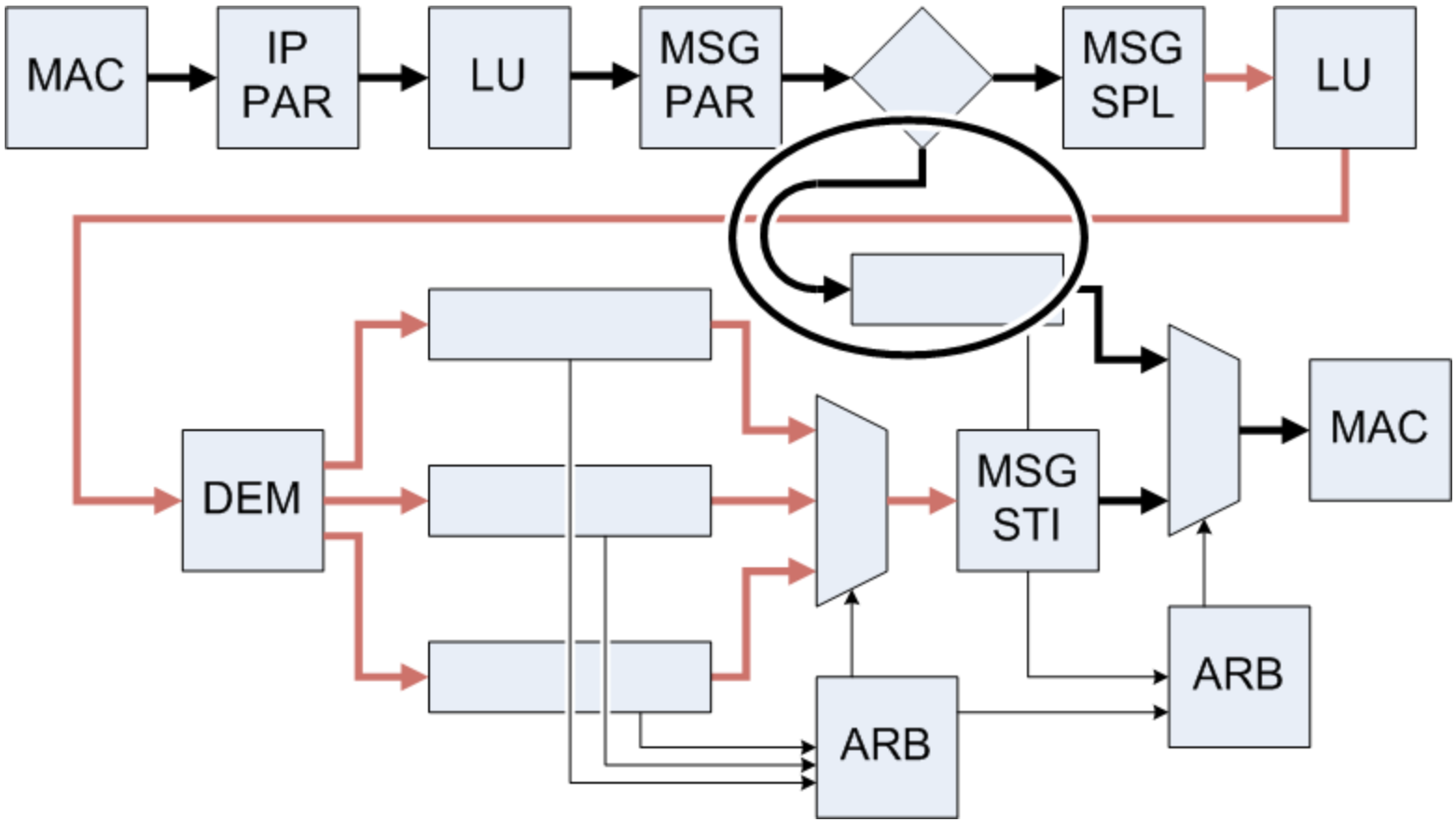
Copyright © 2012 Solarflare Communications, Inc. Slide 32

Parse eligible packets into messages



Copyright © 2012 Solarflare Communications, Inc. Slide 33

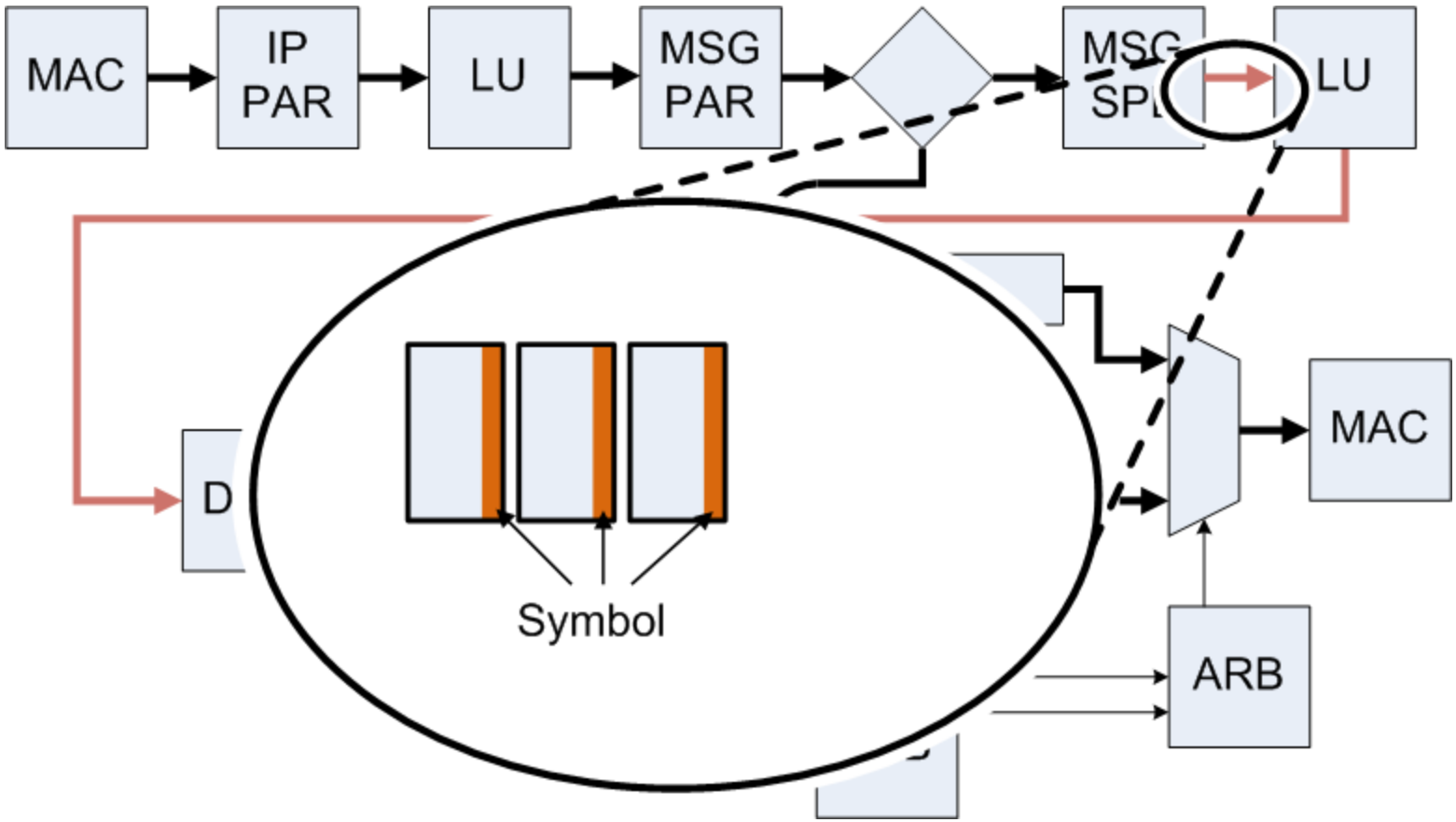
Pass-thru other packets



- Errors during parsing also lead to pass-thru

Copyright © 2012 Solarflare Communications, Inc. Slide 34

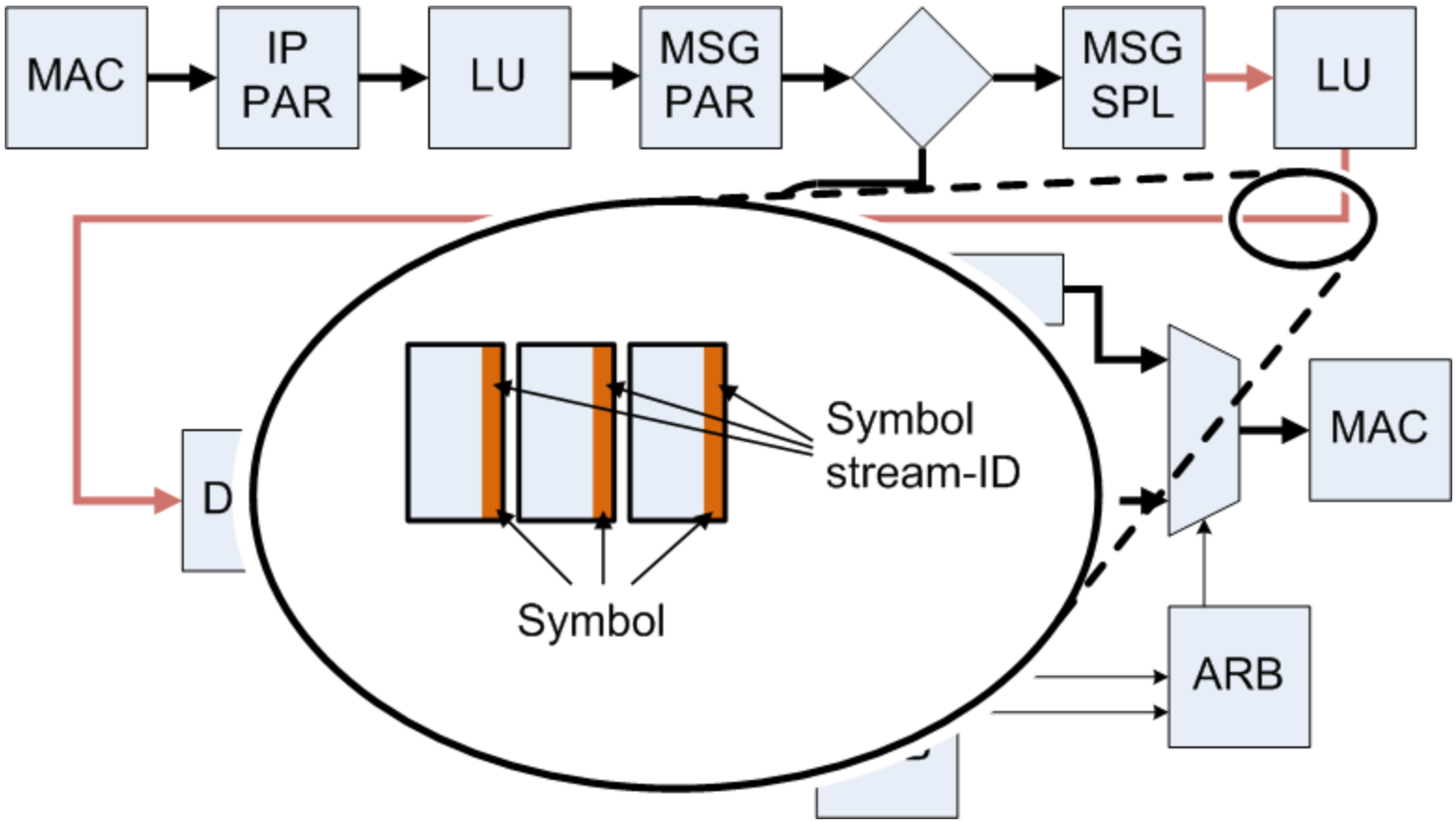
Split packets at message boundaries



- Original headers are discarded at this point

Copyright © 2012 Solarflare Communications, Inc. Slide 35

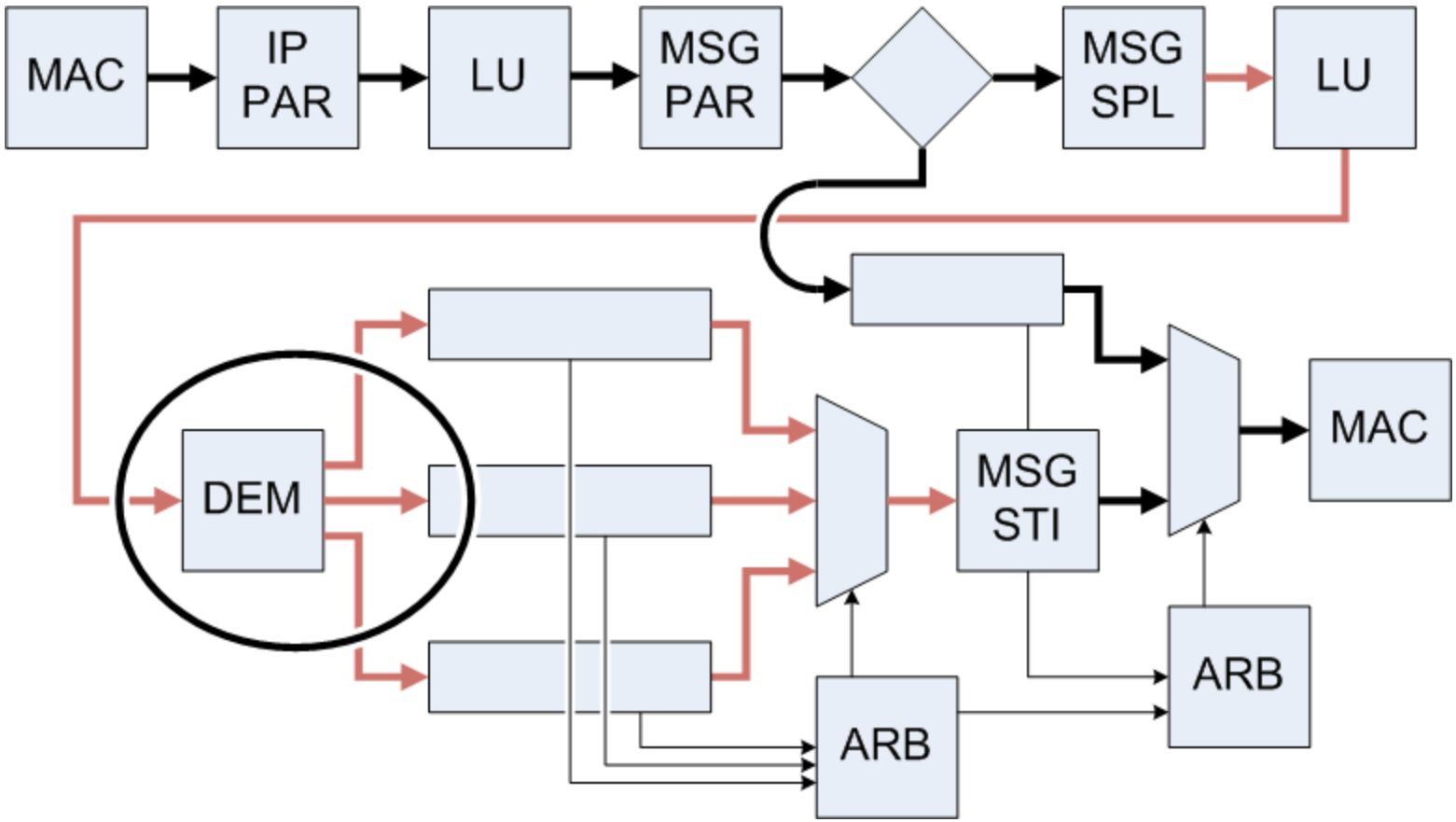
Lookup trading symbol mapping



- Assign integer ID for each output stream

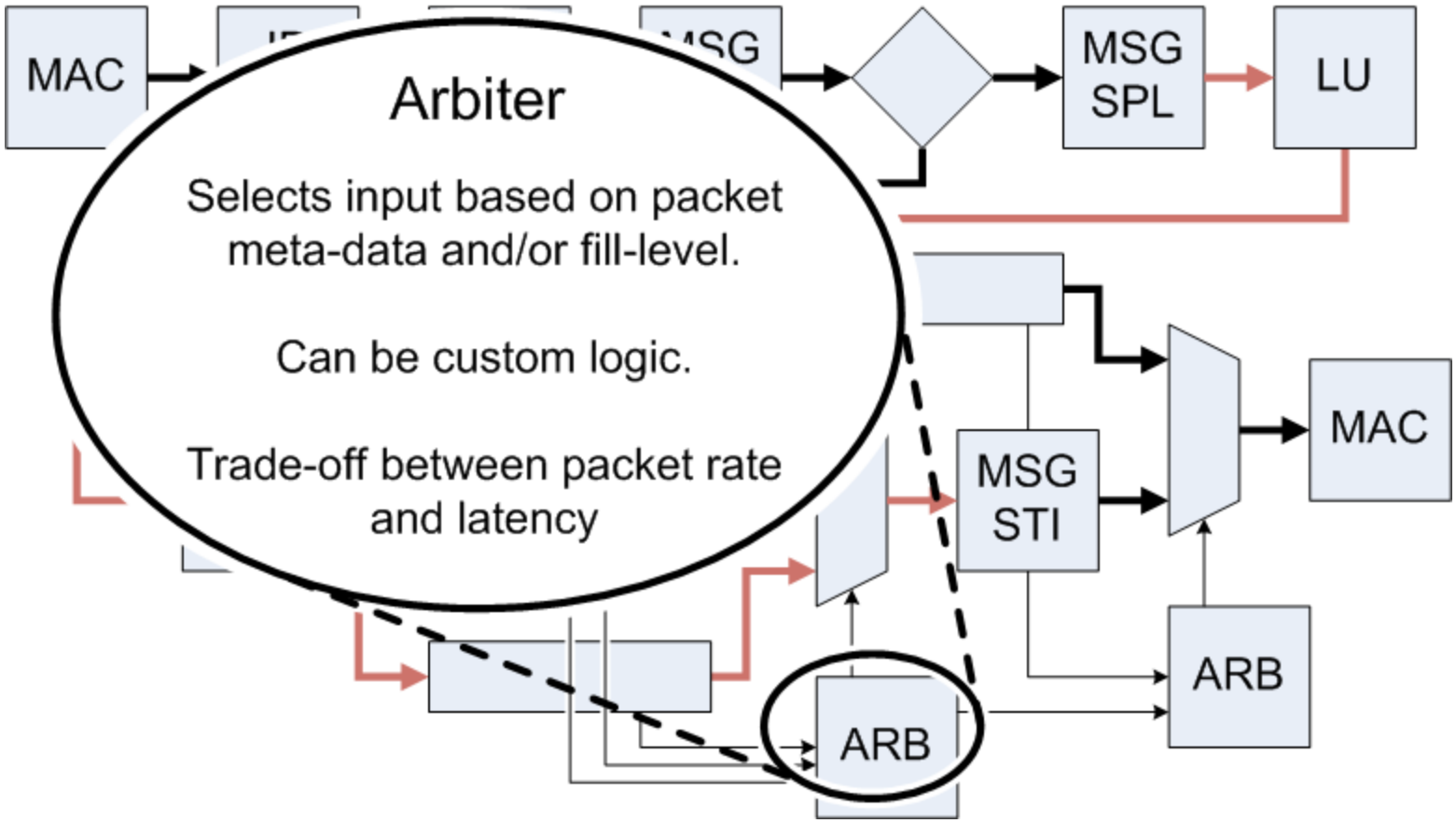
Copyright © 2012 Solarflare Communications, Inc. Slide 36

Symbol stream-ID selects FIFO



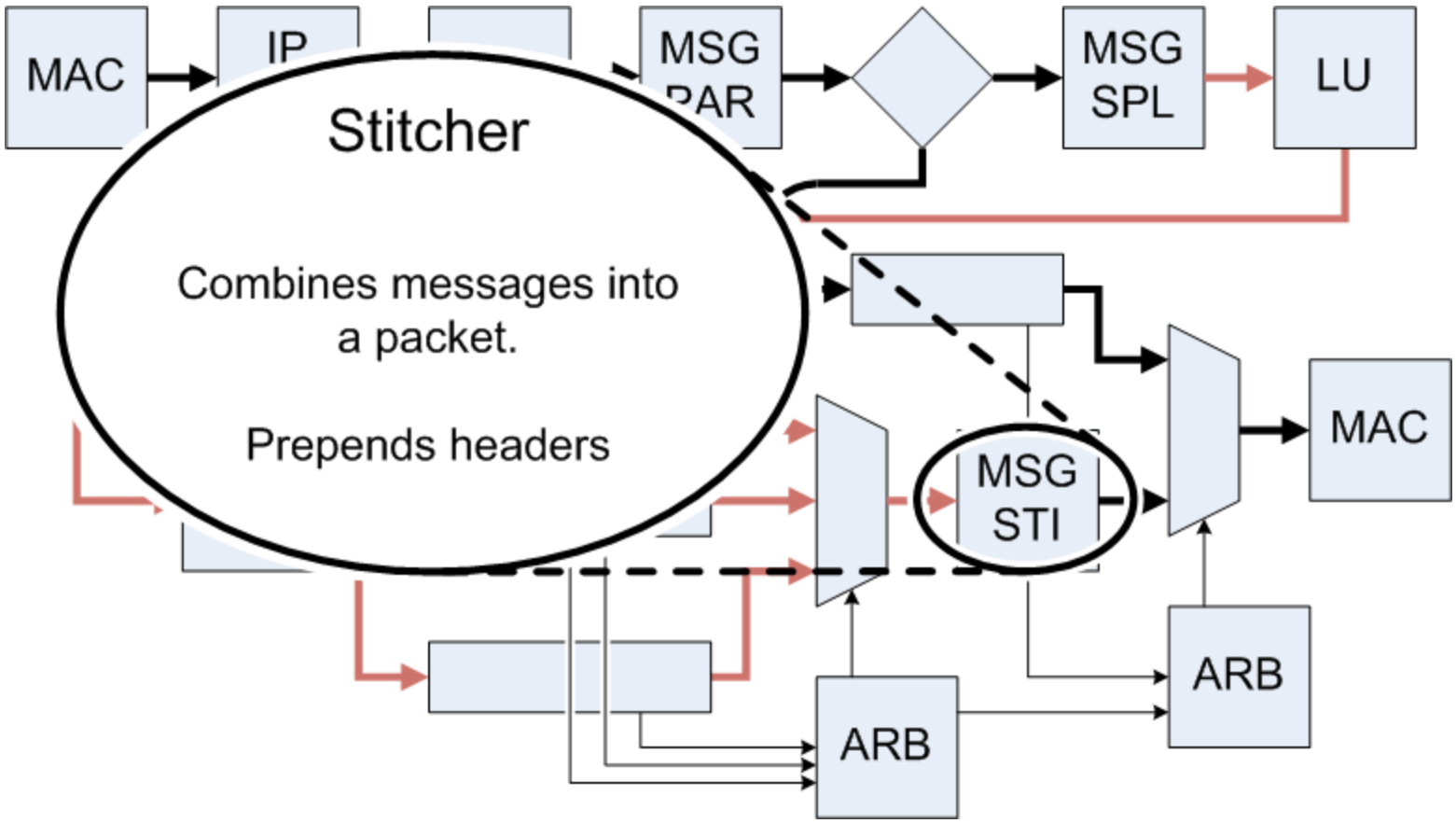
Copyright © 2012 Solarflare Communications, Inc. Slide 37

Arbitrate amongst symbol streams



Copyright © 2012 Solarflare Communications, Inc. Slide 38

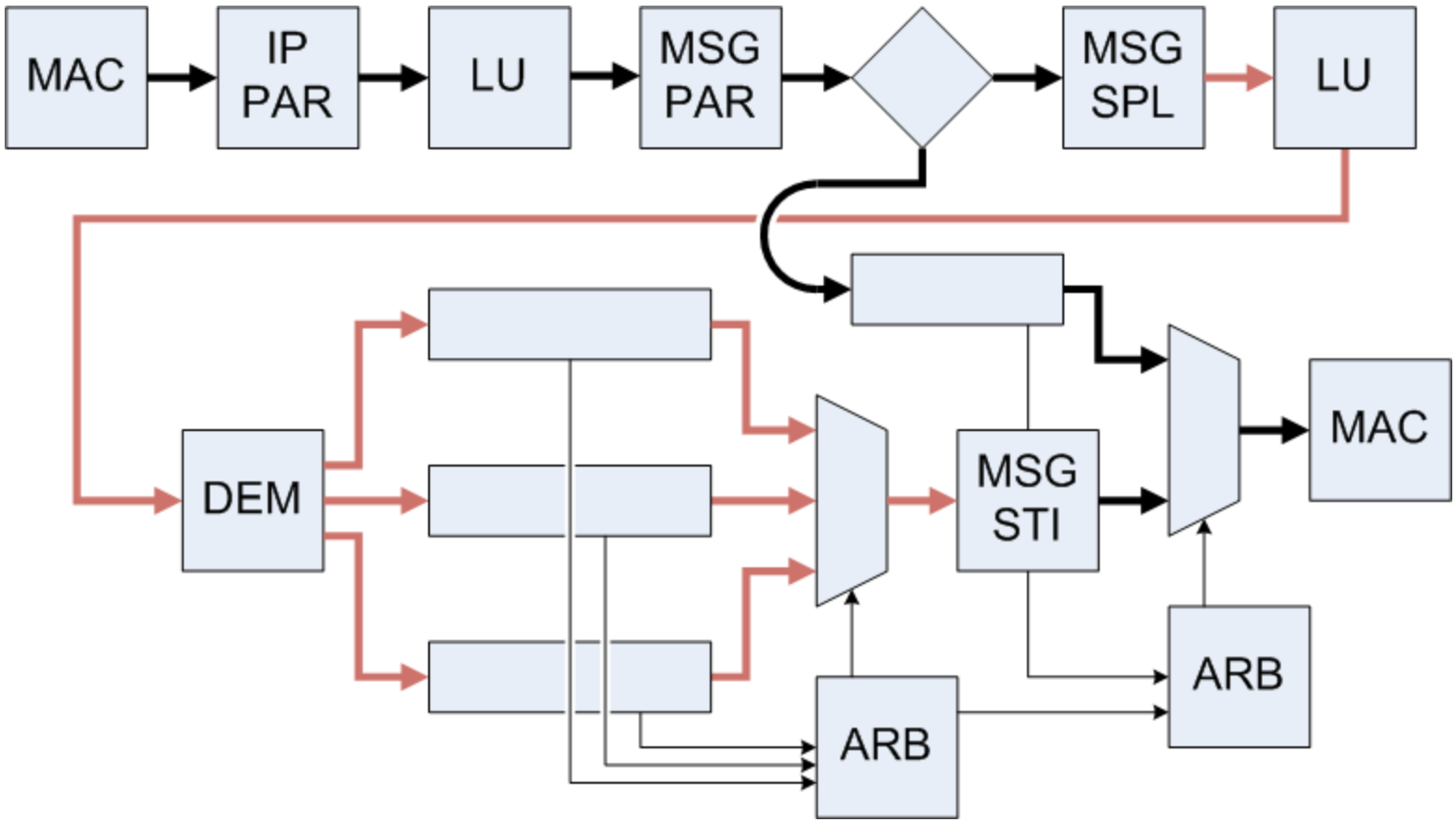
Stitch messages back into packets



- For minimum latency at low rates: One message per packet
- Packet rate limit per stream forces multiple messages per packet

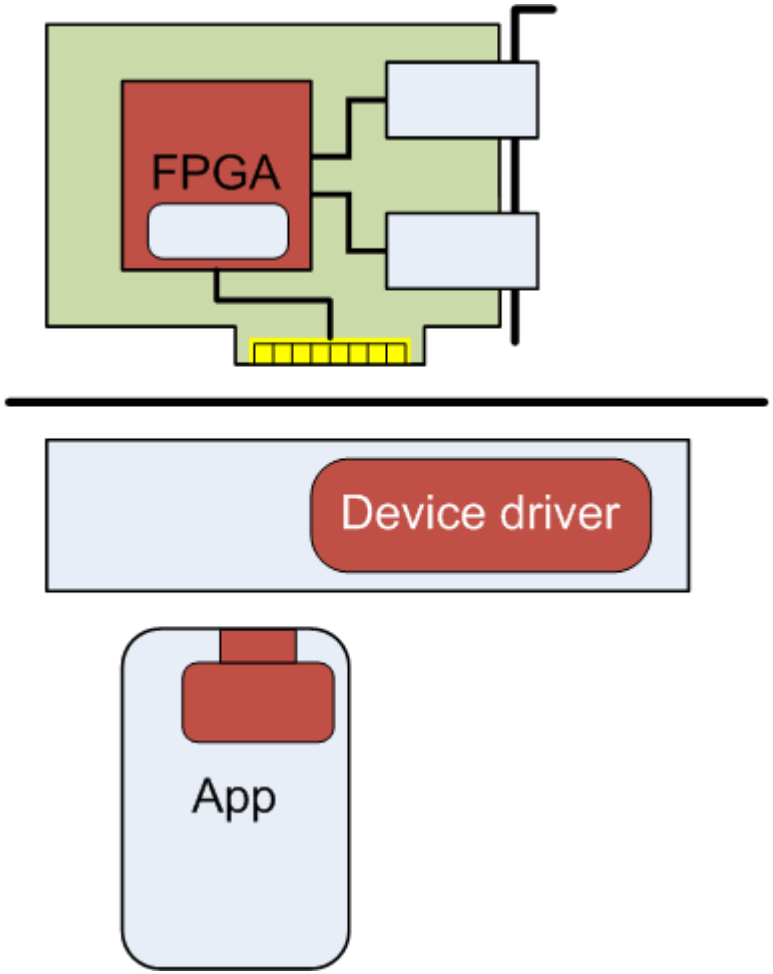
Copyright © 2012 Solarflare Communications, Inc. Slide 39

…and deliver to host



Copyright © 2012 Solarflare Communications, Inc. Slide 40

Custom apps: How much work?

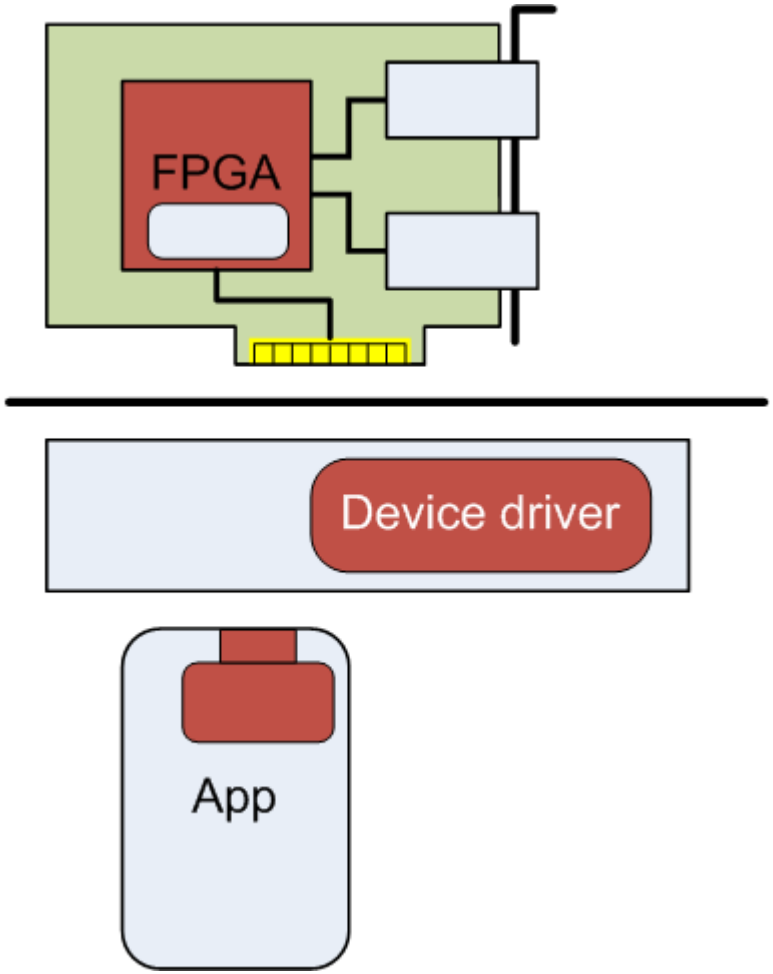


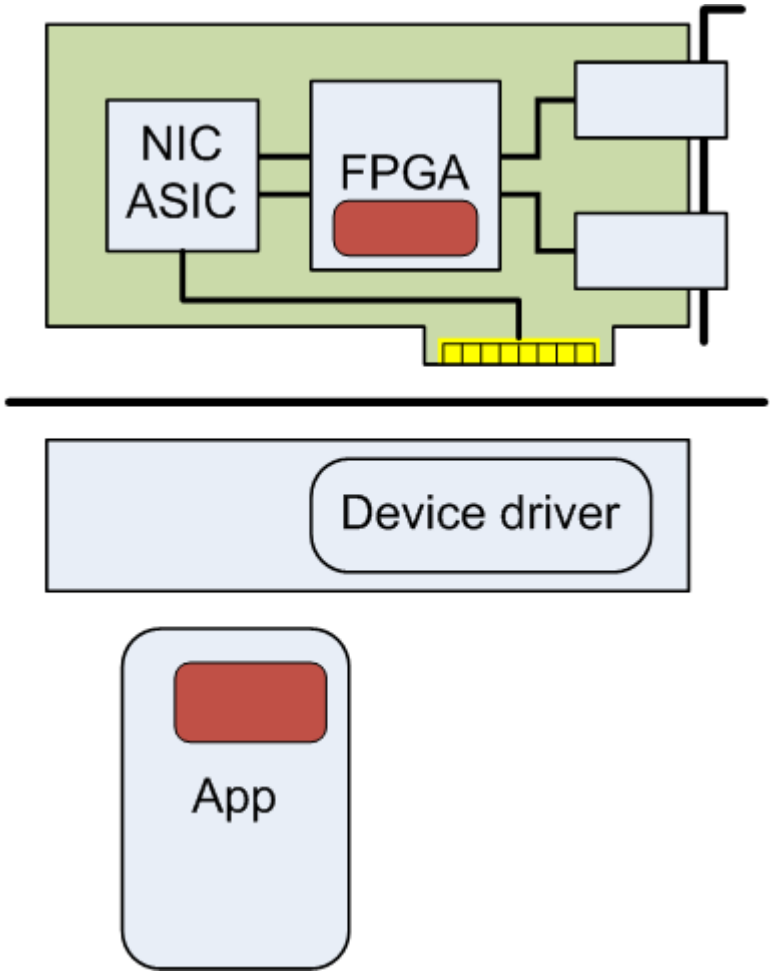
- FPGA image
 (some standard blocks)
- Device drivers
- App/FPGA interface
- App integration

Copyright © 2012 Solarflare Communications, Inc. Slide 41

Custom apps: How much work?

AOE™
ApplicationOnload™ Engine



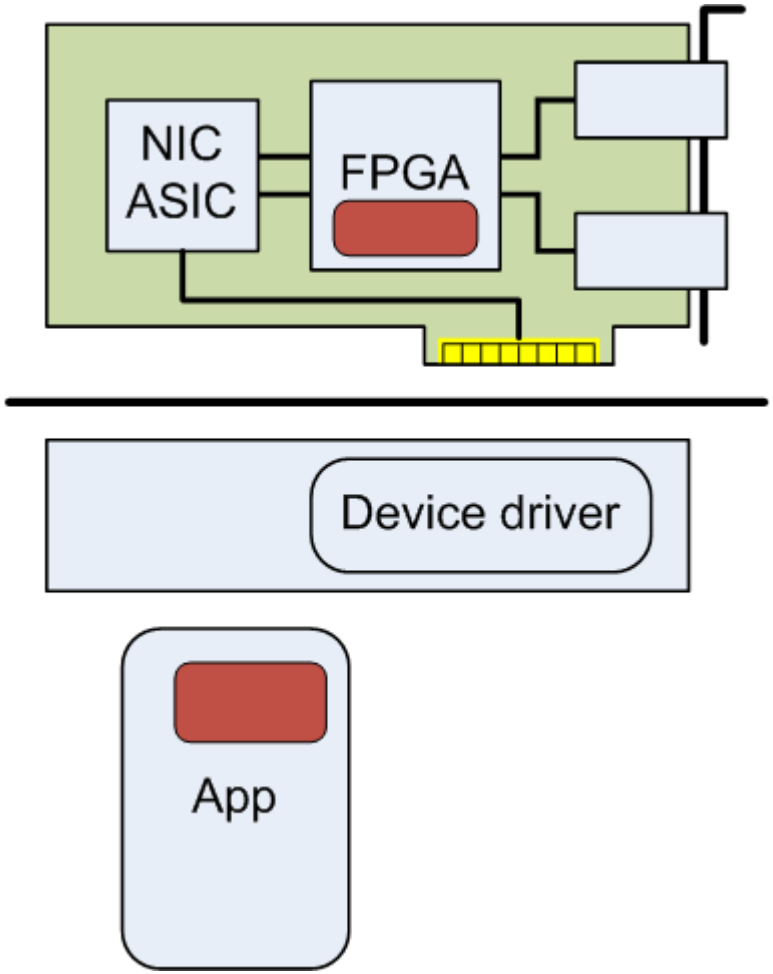


Copyright © 2012 Solarflare Communications, Inc. Slide 42

SOLARFLARE®

Custom apps: How much work?

- FPGA business logic
- App integration



Copyright © 2012 Solarflare Communications, Inc.

Summary

- Solarflare's Application Onload Engine
 - Practical acceleration for network applications
 - Much less work to offload custom business logic
 - Supports incremental deployment
 - Shipping now
 - First deployments being used for enterprise messaging and market data

Copyright © 2012 Solarflare Communications, Inc.

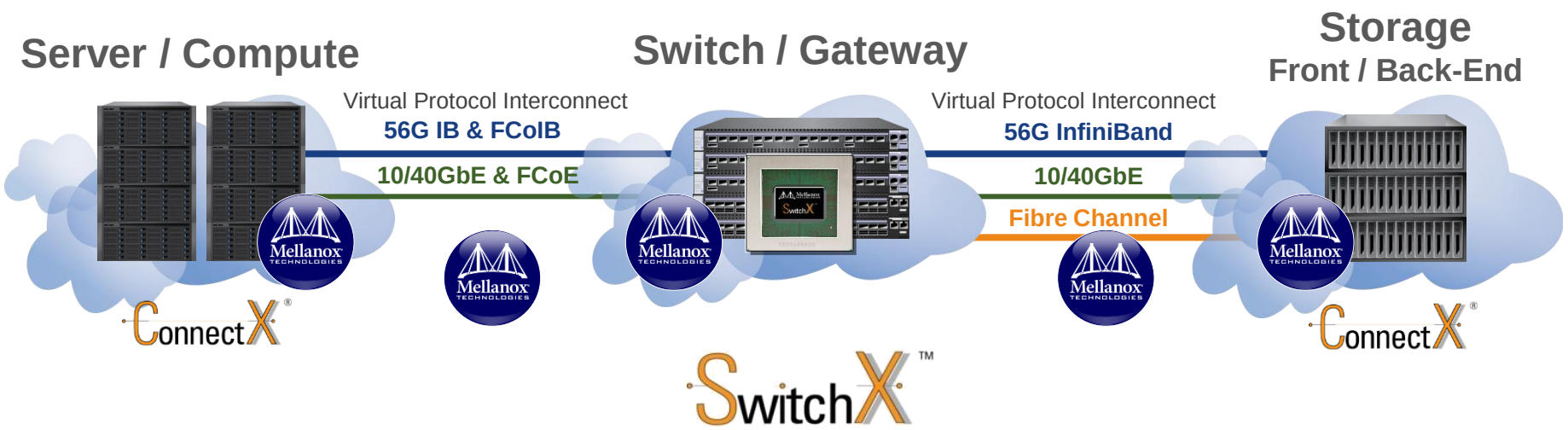


SwitchX

Virtual Protocol Interconnect (VPI)
Switch Architecture

© 2012 MELLANOX TECHNOLOGIES

SwitchX™ - Virtual Protocol Interconnect® Solutions



Mellanox End-to-End Virtual Protocol Interconnect Solution

SwitchX™

- **Fifth generation switching IC from Mellanox**
- **Virtual Protocol Interconnect (VPI) technology – 'One-Wire' fabric for InfiniBand – Ethernet – Fibre Channel traffic**
- **Provides Highest Capacity, Lowest Latency, Lowest Power consumption in the Industry**



© 2012 MELLANOX TECHNOLOGIES

SwitchX™ Performance and Configuration Flexibility

PERFORMANCE

- 4Tb/s
- 36 x 40/56G
- 200ns Latency
- 40 Watts @ 64 10GE
- 55 Watts @ 36 40GE

1U switch configuration options

- 36 Port FDR IB
- 36 Port 40GigE VPI IB/Ethernet
- 64 Port 10GigE VPI IB/Ethernet
- 12 Port 40GigE/48 Port 40GigE VPI

Blade switch configuration options

- 16 - 40GigE to servers
- 12 - 10GigE to LAN
- 8G FC to SAN/2 - 40GigE stacking ports

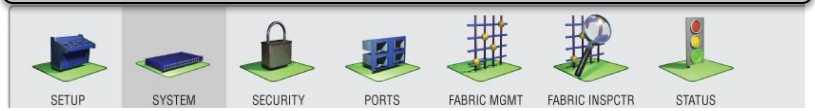
Modular switch chassis options

- Up to 648 56G IB ports
- Up to 648 40GigE ports

SwitchX™ VPI Switch

Unified Fabric Manager

Switch OS Layer





- 64 ports 10GbE
- 36 ports 40GbE
- 48 10GbE + 12 40GbE
- 36 ports IB up to 56Gb/s
- 8 VPI subnets

1U switches

Blade switches

Modular switches



© 2012 MELLANOX TECHNOLOGIES

Ethernet Switching Performance

- Throughput (2.5X)
 - 2.88Tb/s throughput on a single chip, running Full Wire Speed at any packet size
- L2 UC/MC Latency for L2/L3 switches (2X)
 - 198-223ns for any packet size
- L3 Latency (2X)
 - 321-337ns for any packet size
- Power Efficiency (6X)
 - Sub 0.6Watt per 10GbE throughput with 100% load at Full Wire Speed







© 2012 MELLANOX TECHNOLOGIES

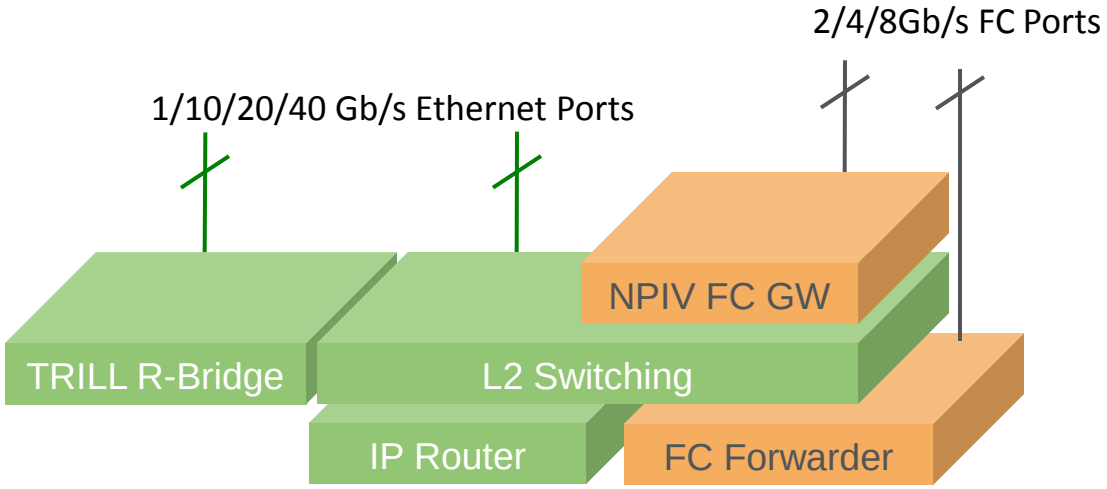
Feature Overview

Physical Layer	SFI XAUI/RXAUI/XLAUI KR/KR2/KR4 - 10/20/40/56 GigE IB SDR, DDR, QDR, FDR10, FDR - 10/20/40/56/100 Gb/s 2/4/8 Gb/s Fibre Channel
L2/L3 Protocols	Ethernet, Datacenter Bridging, TRILL , QCN, Fibre Channel Forwarder, IP Routing, IB Routing
Bridges	Ethernet to Fibre Channel Gateways (NPIV) IB-Ethernet and IB-FC Gateways
QoS	802.1p and DIFFSERV classification, marking Access Control Lists (iACL, eACL)
Management	Dual SGMII PCIe Gen2 x4, or PCIe x1 and Dual SGMII GPIOs, I2C, JTAG
Other Features	Energy Efficient Ethernet, Active Power Governor IEEE 1588 time stamping 40nm process, 1924 pins in 45x45mm FCBGA 1 mm pitch

© 2012 MELLANOX TECHNOLOGIES

Converged I/O Solutions



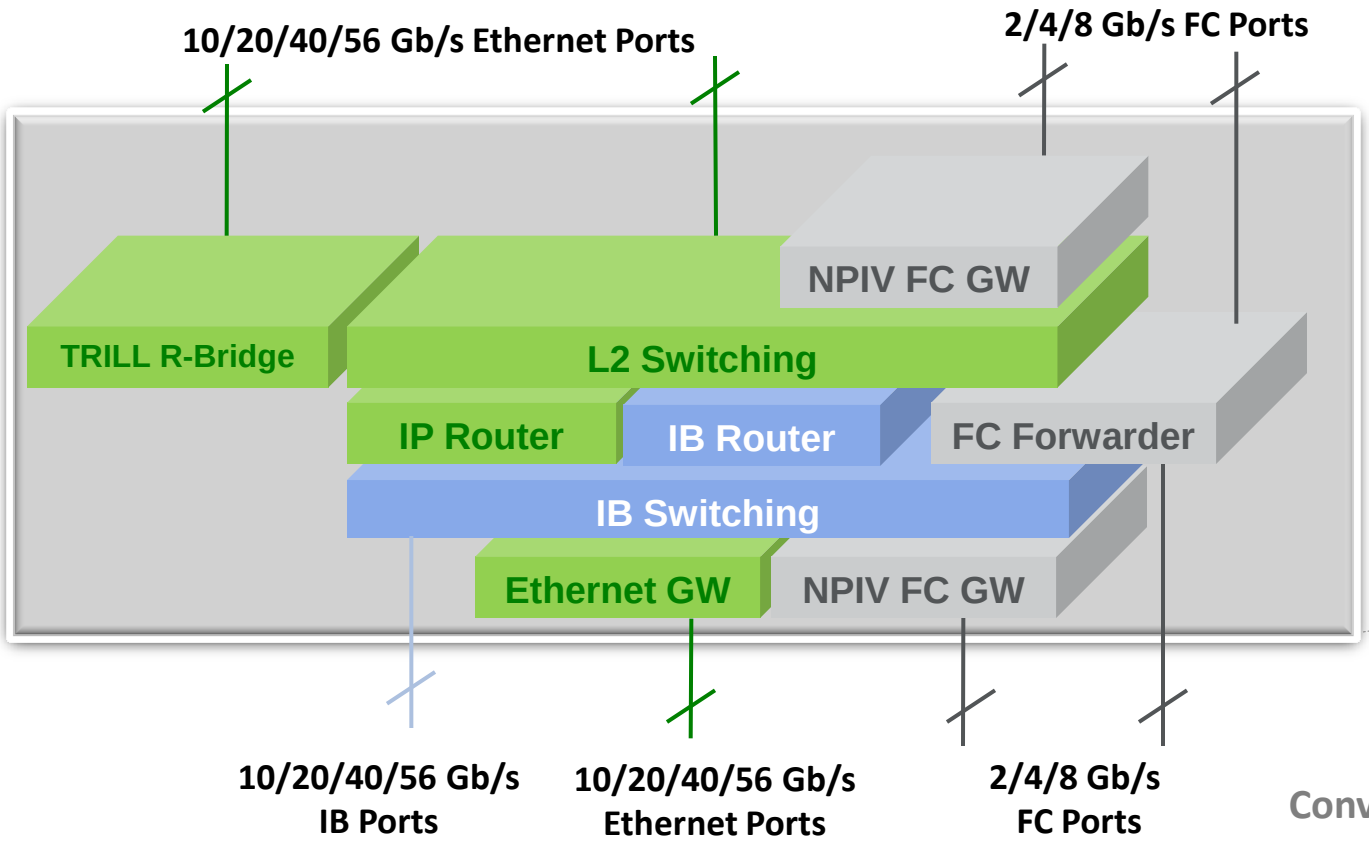


- 2.8/4 Tb/s lossless switching
- 64 10GigE, 36 40GigE
- Flexible mix of ports
 - i.e. 48 10GigE, 12 40GigE
- Multi-chip, high port count configurations
 - Efficient cluster scaling
 - Fat tree scaling
 - Adaptive routing

- NPIV, FCF based native FC ports
 - 2/4/8 Gb/s
 - N, VN, F, VF, E, VE port types
 - Soft and hard zoning
- Sample port configuration
 - 40 10GigE, 24 8Gb/s FC
 - 52 10GigE, 12 8Gb/s FC
 - 24 40GigE, 24 8Gb/s FC
 - 30 40GigE, 12 8Gb/s FC

© 2012 MELLANOX TECHNOLOGIES

Virtual Protocol Interconnect (VPI) IO Convergence





Converged I/O with Ethernet

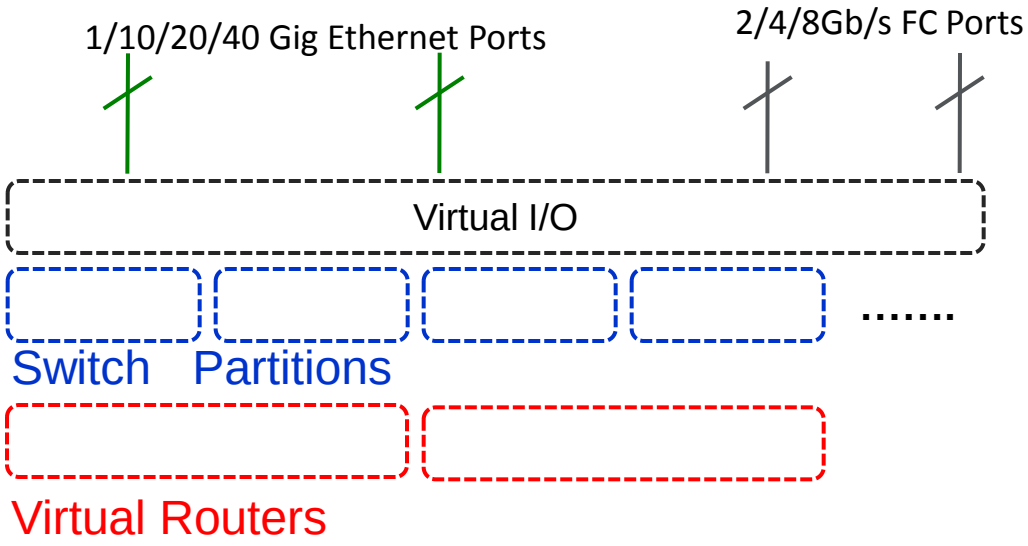
Repurposing

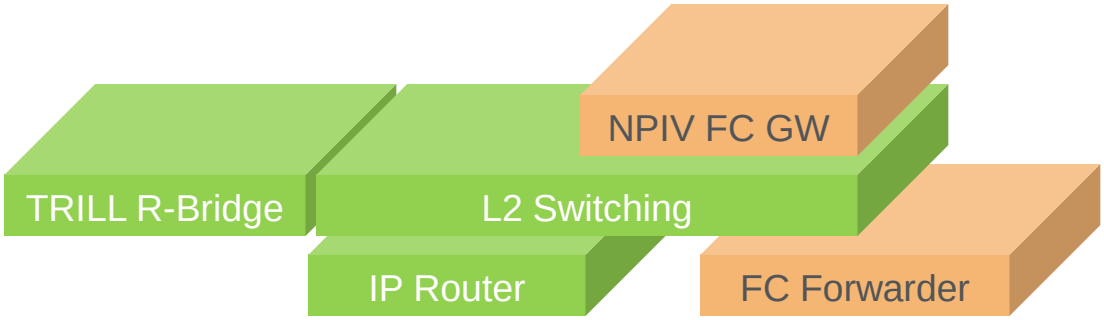
Converged I/O with InfiniBand

© 2012 MELLANOX TECHNOLOGIES

Switch Partitioning

- Up to 8 switch partitions can be activated
- Flexible number of Ethernet, FCoE, FC port assignments per switch partition
- Separate L2 data and control plane domains
- Multiple Virtual Routers
- Separate address space per VR
- Isolation and fault containment





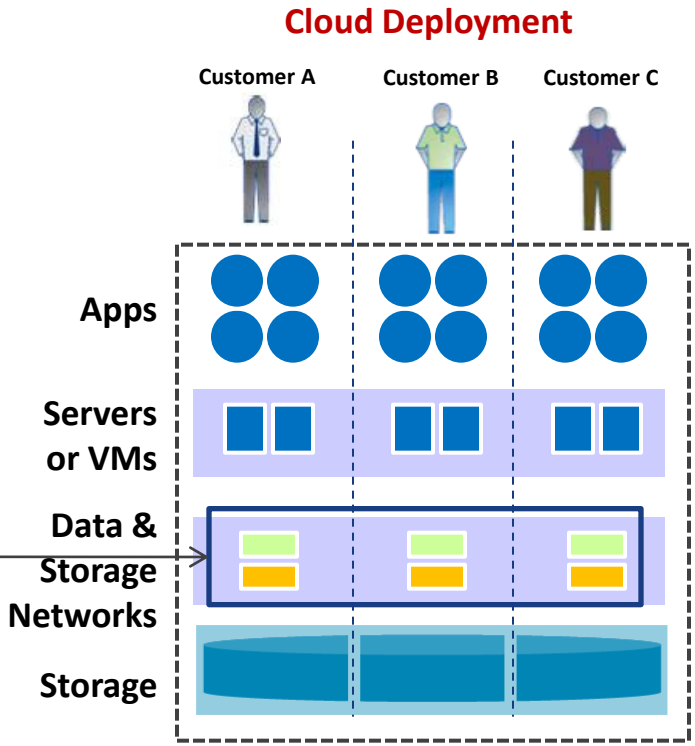
© 2012 MELLANOX TECHNOLOGIES

Control Plane Multitenancy

- Multiple switch partitions can be instantiated
 - Like virtual switches inside physical switch
 - Complements virtualized servers and storage
 - Control/data separation like separate switches
- Flexible # of ports & personalities
 - Per switch partition, e.g., IB, L2+ Eth, FC

SwitchX™

With Switch Partitions

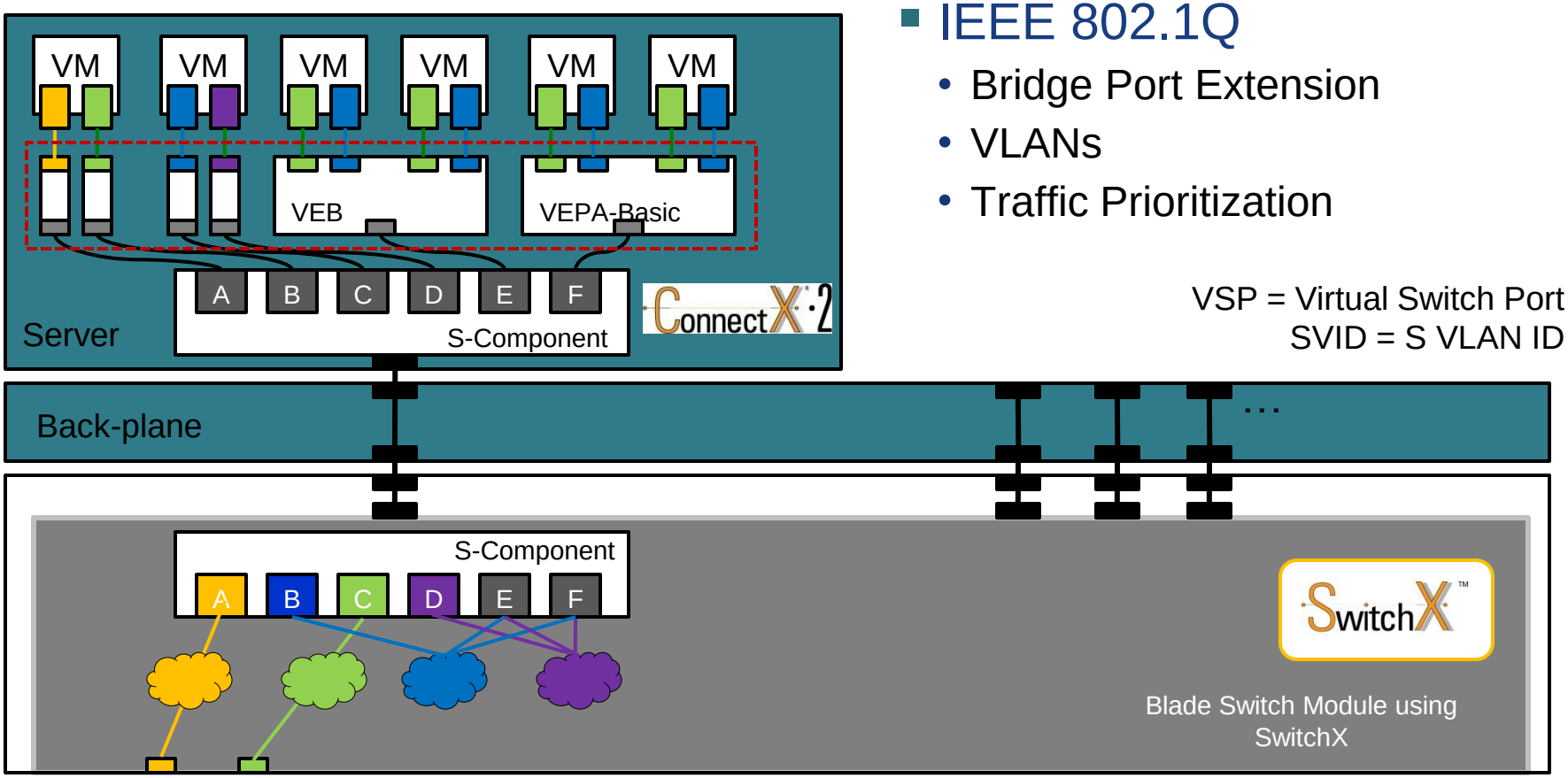


Supports evolving cloud & multi-tenancy architectures

© 2012 MELLANOX TECHNOLOGIES

Switch Port Virtualization

- Flexible VSP Allocation
 - 16 VSPs on 18 ports
 - 8 VSPs on 36 ports
 - 4 VSPs on 64 ports
- Hairpin Mode per VSP
- Switch Partition per VSP
- SVID Allocation
- IEEE 802.1Q
 - Bridge Port Extension
 - VLANs
 - Traffic Prioritization



VSP = Virtual Switch Port
SVID = S VLAN ID

© 2012 MELLANOX TECHNOLOGIES

SwitchX Multi-chip Stacking

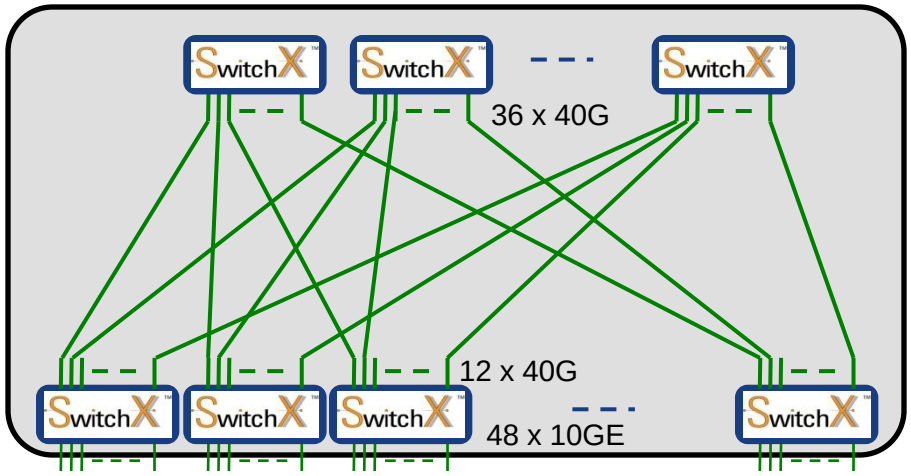
- Chain and ring topologies
- Any port can be a stacking port
- Single point of management across stacking units (SU)
 - Efficient Inband configuration over management datagrams (eMAD)
- System resiliency
 - Any SU can take charge of the system
 - Alternate paths dynamically used when stacking link down
- Cross system features
 - Link aggregation – ports across SUs in same LAG group
 - ACL – same policy to ports across SUs
 - e.g. VLAN ACL
 - Unified tables are populated on all SUs
 - e.g. L2 filtering DB, L3 routing tables



© 2012 MELLANOX TECHNOLOGIES

Multi-chip Configuration - Topologies

2 Layers FAT-TREE

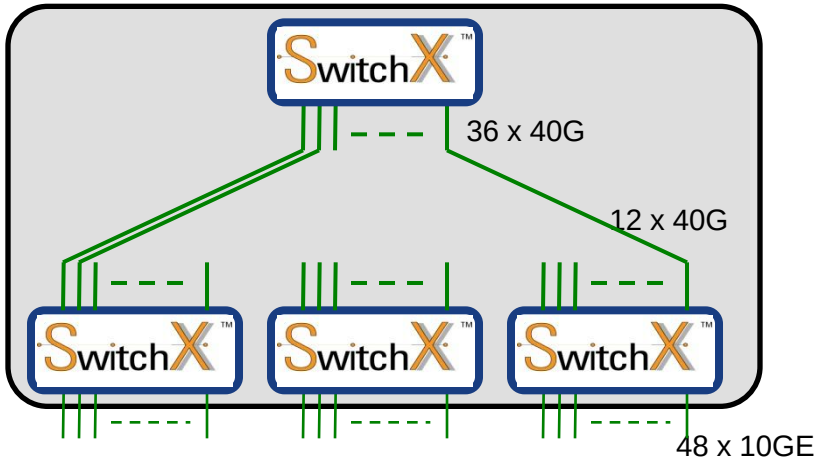


Up to 1728 10GE Ports



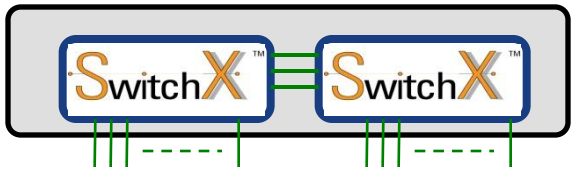
2 Layers FAT-TREE

-Single Spine Chip



Up to 144 10GE Ports

Back to Back



Up to 96 10GE Ports

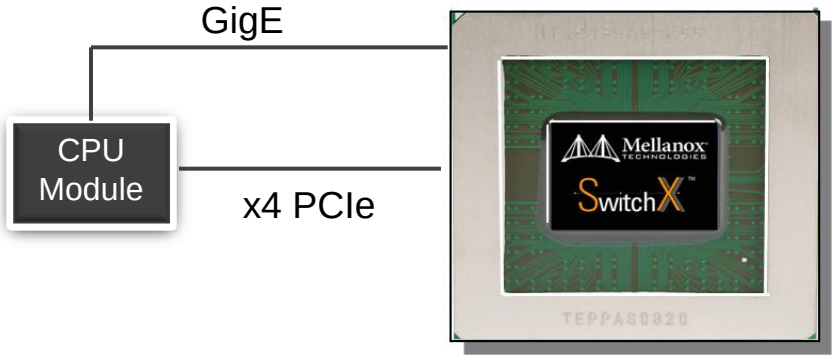
- Non Blocking
- L2 and L2 Multicast forwarding
- Link Aggregation across fabric
- Port Mirroring across fabric
- Seamless class of service support
- Preserving VLAN membership

This slide does not present all possible configurations – but rather most reasonable multi-chip configuration topologies

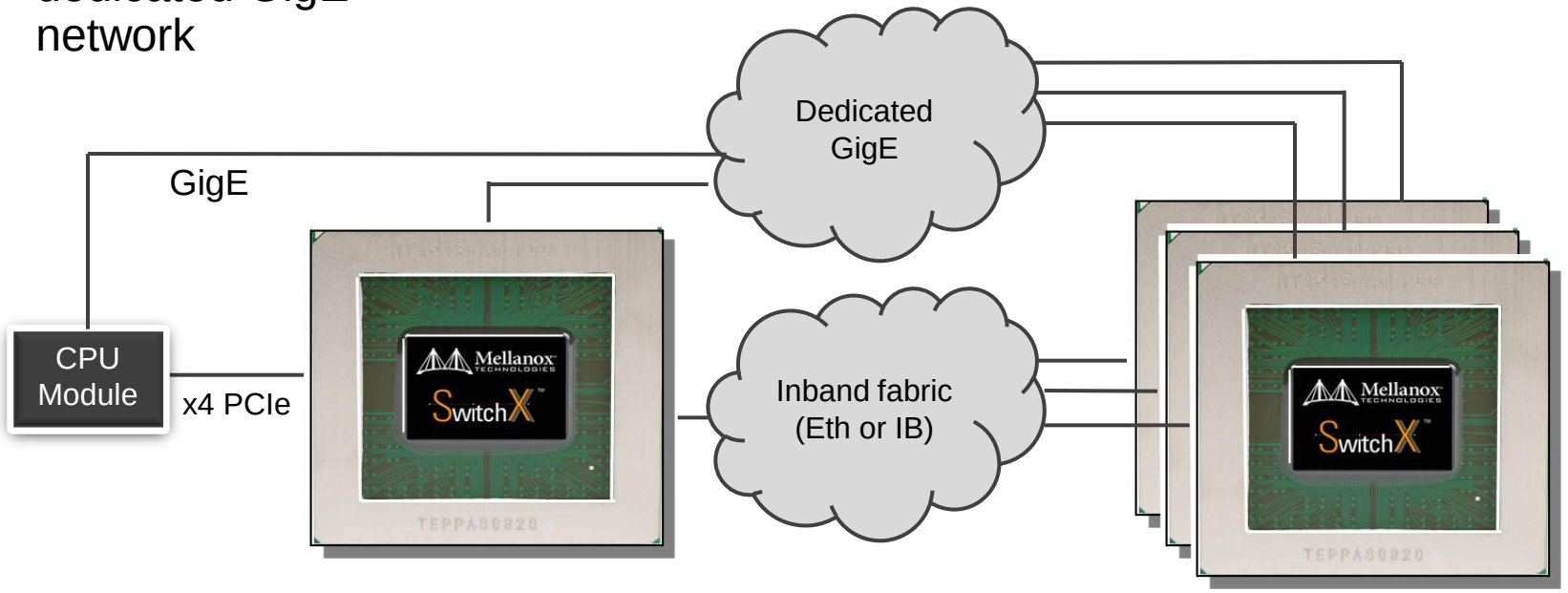
© 2012 MELLANOX TECHNOLOGIES

Flexible Switch Management Interface

- Single chip
 - x4 PCIe or dedicated GigE network



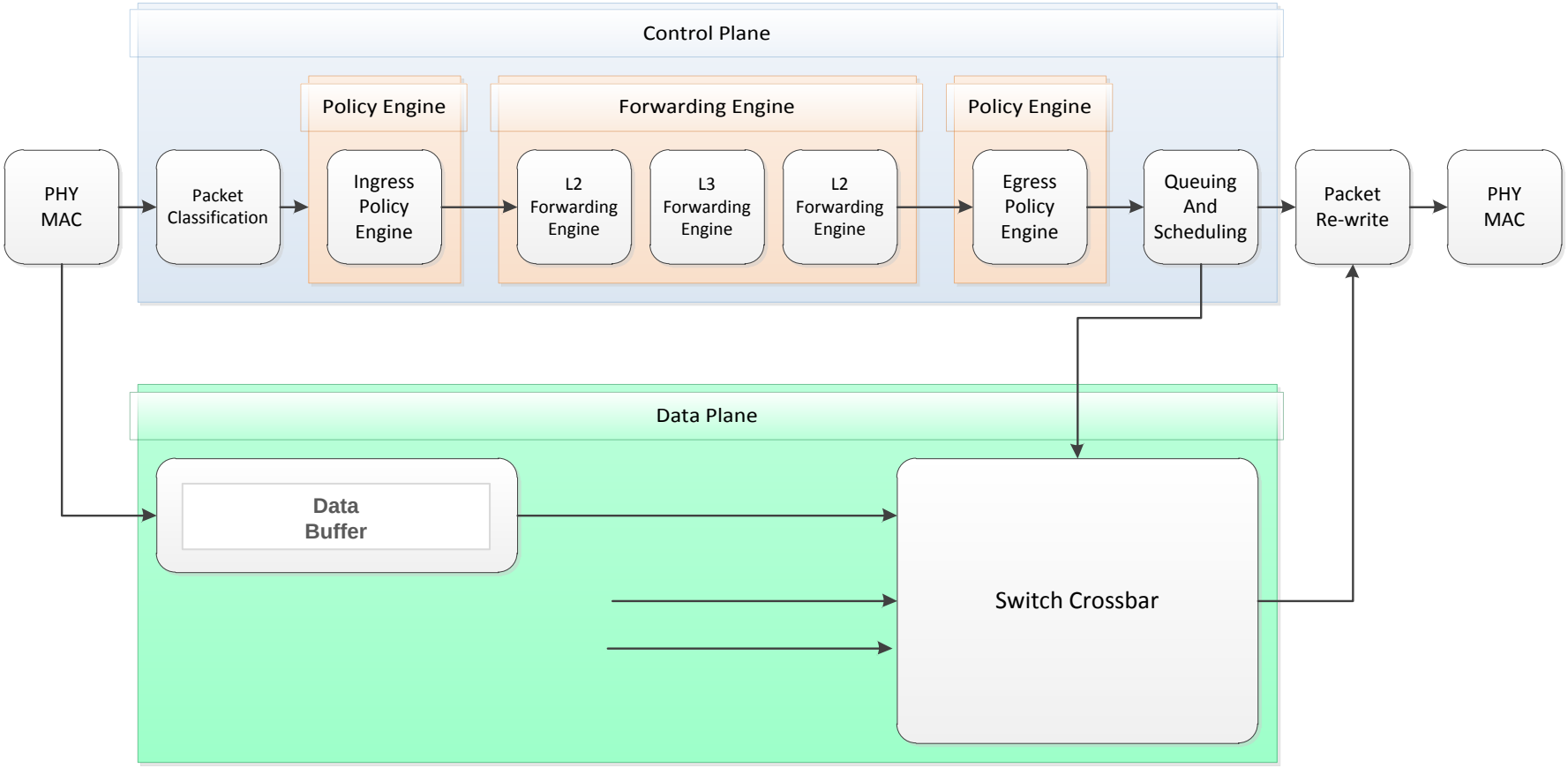
- Multi chip
 - x4 PCIe to inband fabric (Eth or IB) or dedicated GigE network



© 2012 MELLANOX TECHNOLOGIES

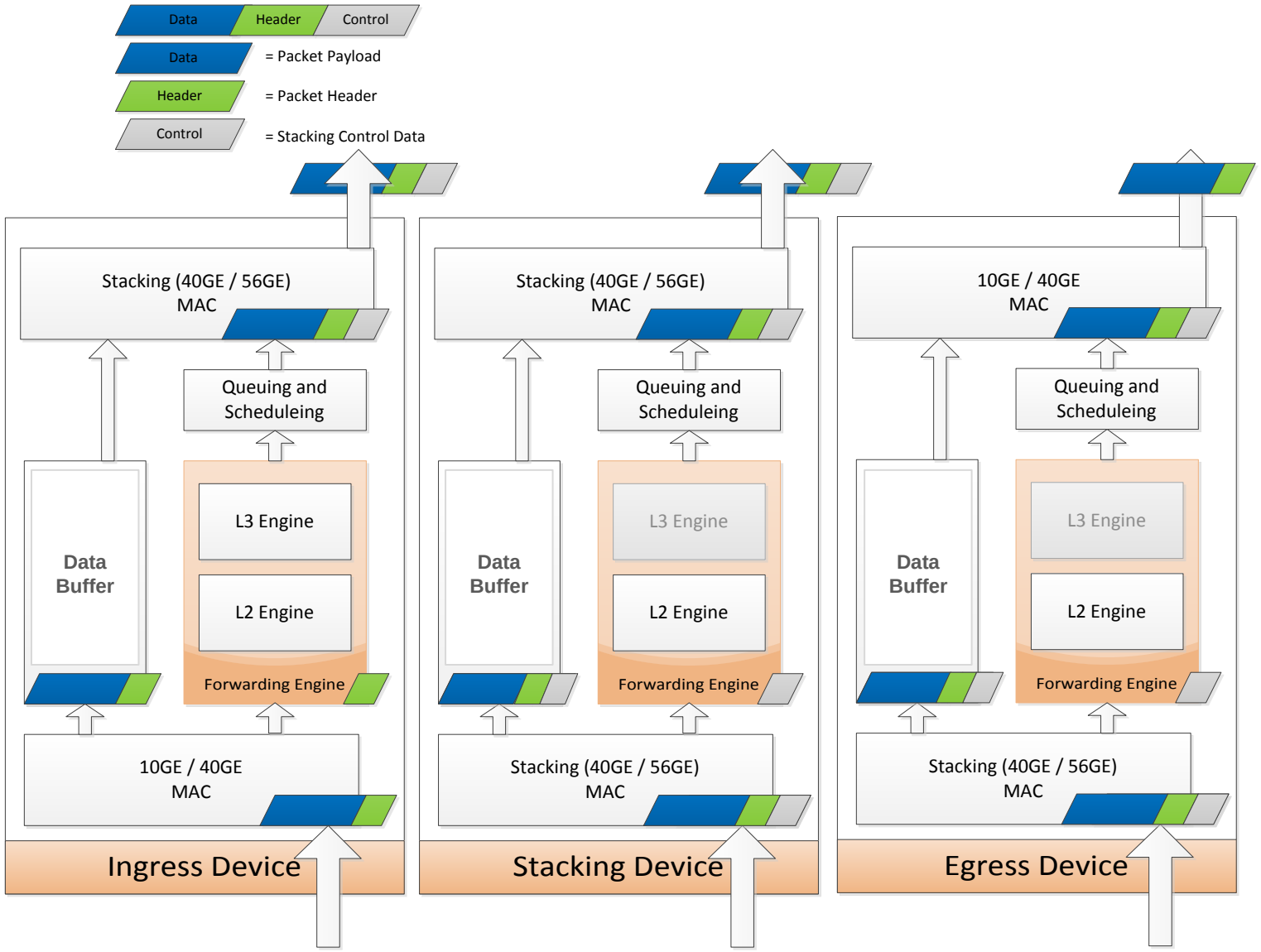
SwitchX Packet Flow Overview

- Fully Pipelined Implementation
- Low-latency cut-through switching support
- Wire speed forwarding



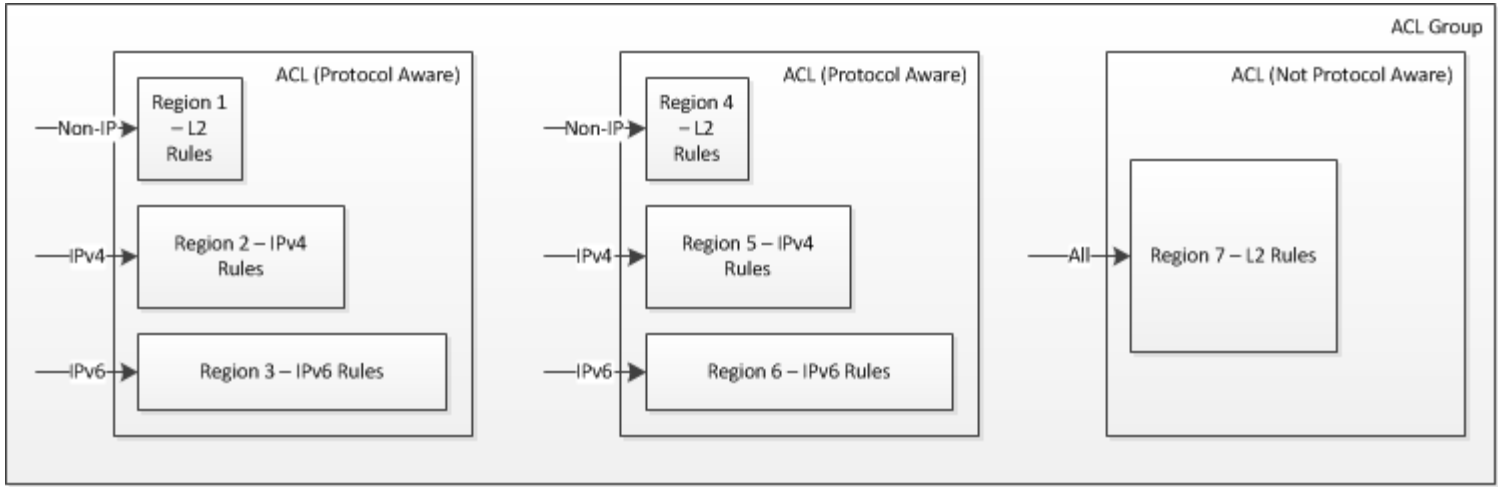
© 2012 MELLANOX TECHNOLOGIES

Forwarding Engine – Packet Flows



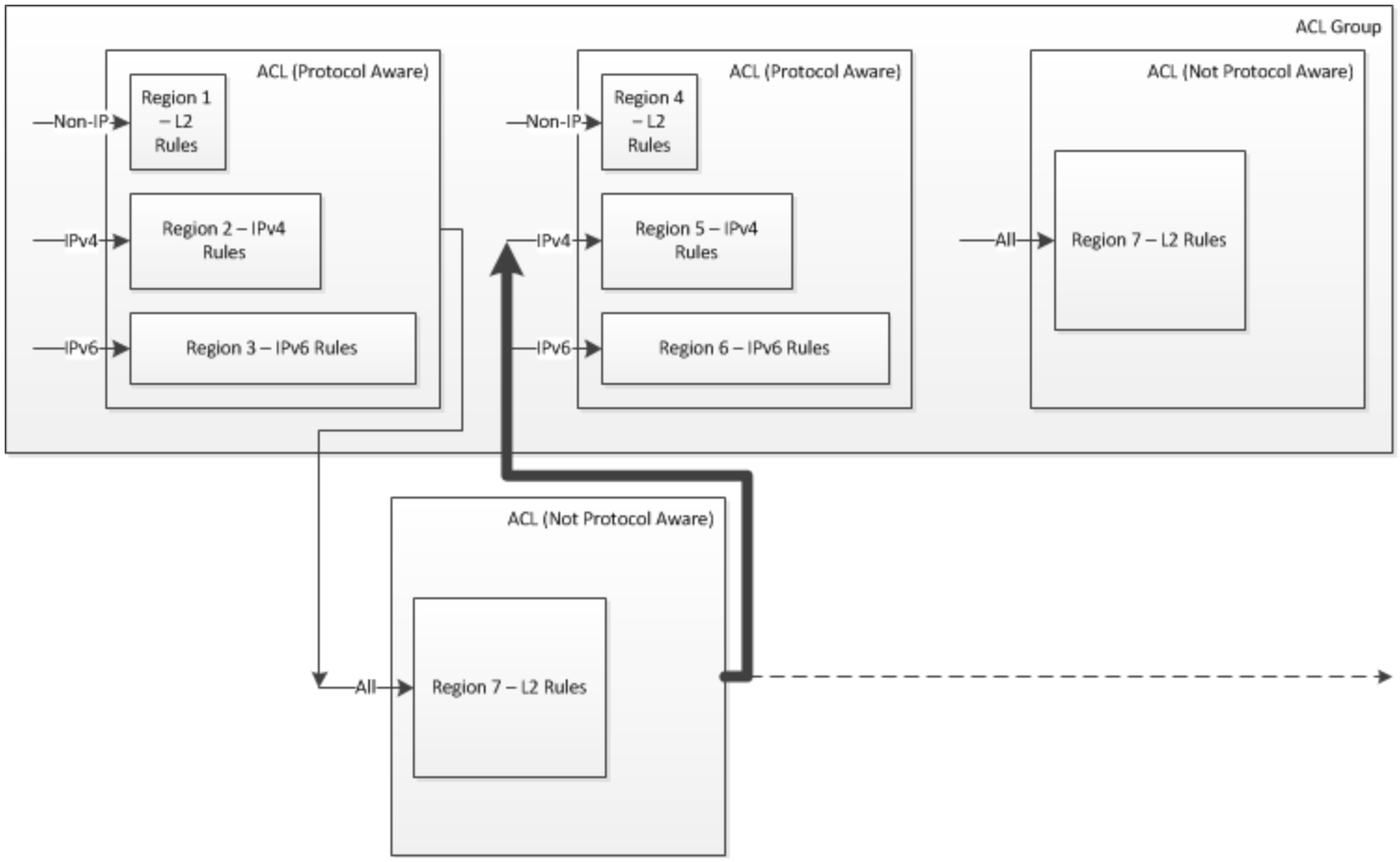
© 2012 MELLANOX TECHNOLOGIES

ACLs Architecture Overview



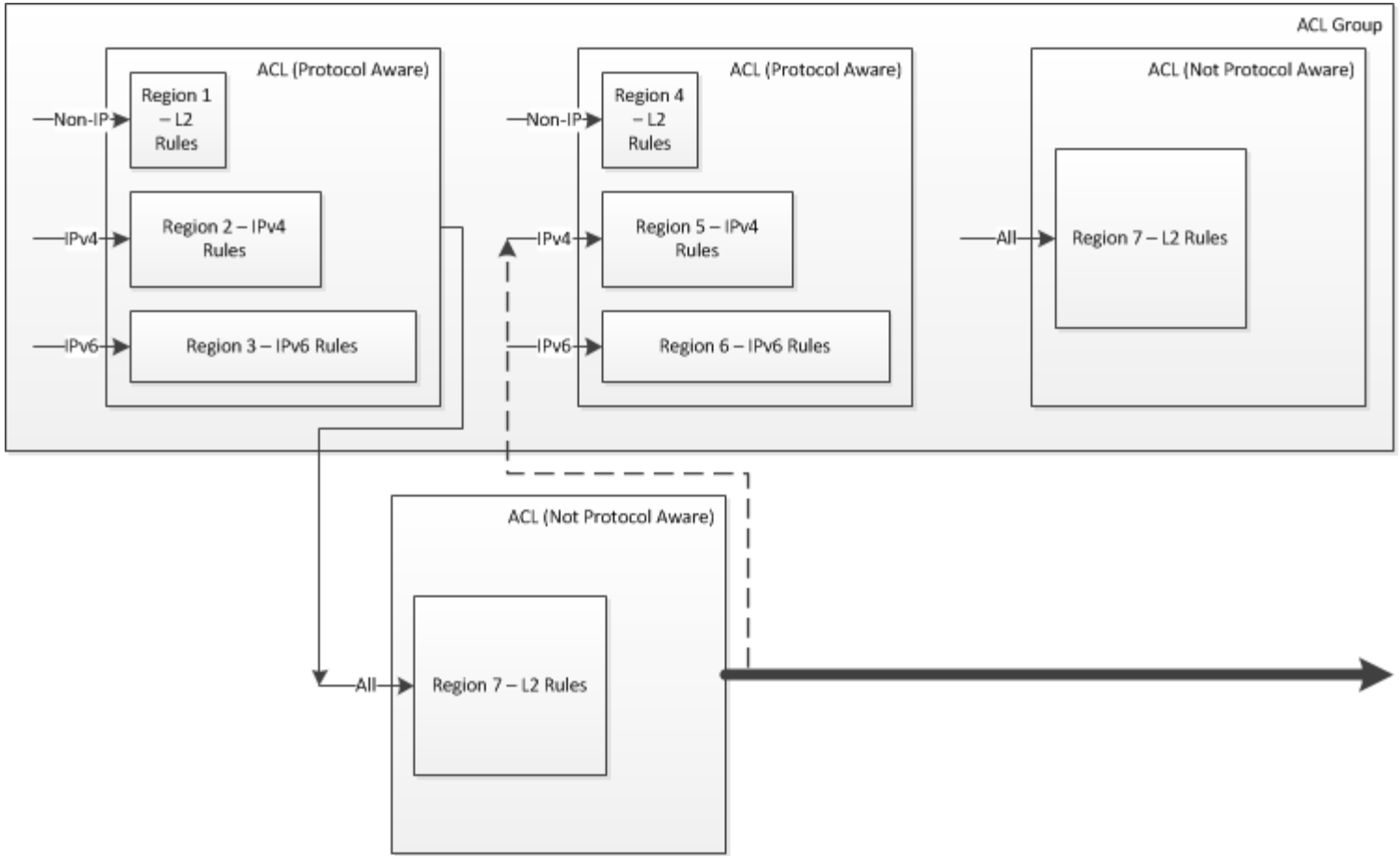
© 2012 MELLANOX TECHNOLOGIES

Rule Binding - Nesting



© 2012 MELLANOX TECHNOLOGIES

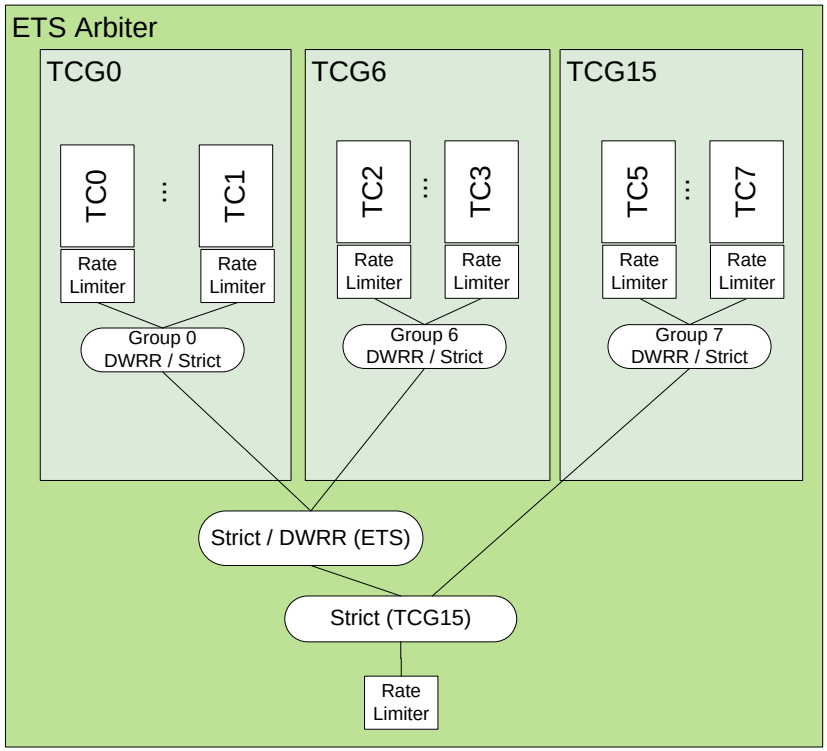
Rule Binding - Break

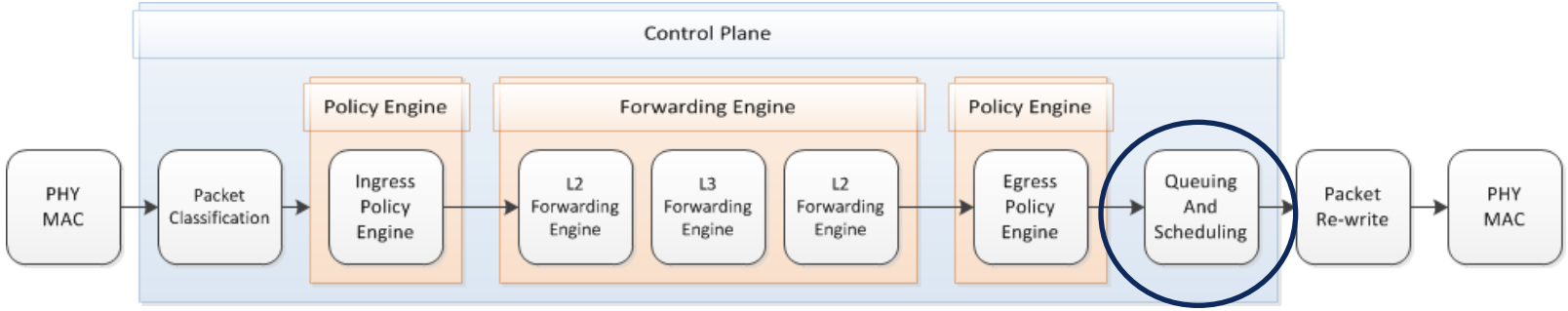


© 2012 MELLANOX TECHNOLOGIES

Queuing and Scheduling

- 8 Traffic Classes
- ETS Scheduling
- Mirroring/Replication
- UC/MC Flows





© 2012 MELLANOX TECHNOLOGIES

SwitchX TOR Portfolio

Ideal as a ToR/Core using efficient 40GbE links between 1st and 2nd tiers



SX1036

ToR with 960G of BW equally split between 10G downlinks and 40G uplinks



SX1024

Ideal as a ToR connected to a 3rd party core switch with no 40GigE links



SX1016

- Capacity
 - 36 40GbE ports
 - 64 10GbE ports
 - 48x10GbE+12x40GbE combo
 - Various other port schemes via breakout cables
- Key Features
 - L2/L3 stack
 - VPI
 - 56GbE
 - End to end solution
- Latency
 - 220ns latency 40GbE
 - 330ns L3 latency
 - 270ns latency 10GbE
 - 430ns L3 latency
- Throughput
 - 2.88Tb/s of non-blocking throughput
- Power
 - Under 1W per 10GbE interface
 - 2.3W per 40GbE interface
 - 0.6W per 10GbE of throughput

© 2012 MELLANOX TECHNOLOGIES

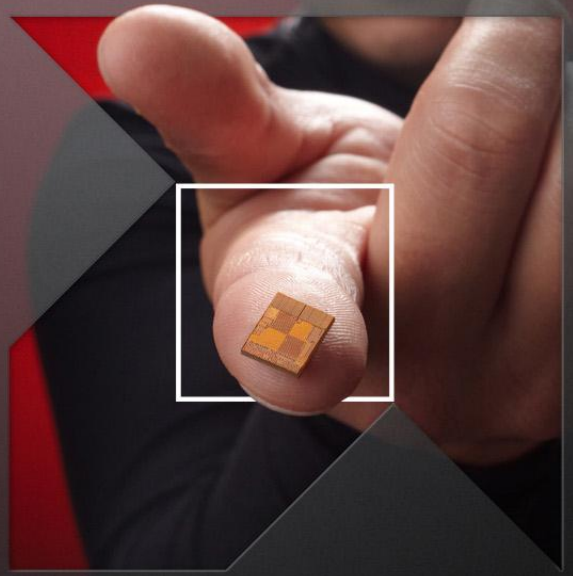
Mellanox SX6536

- 648 x QSFP 40GE* ports
- 1152 x SFP+ 10GE* ports
- 51.84Tb/s throughput
- 9.6 Watt/40GE port
- Latency: 700ns inter line, 230ns same line
- World's first Cut-Through modular Ethernet switch
- N+N PS Redundancy
- L2/L3 SW Stack
- Same Chassis is used for IB FDR (56Gbps)
- Smaller Chassis (324p, 216p, 108p)
 - Same leafs, spines, management boards
 - Same architecture



Thanks





THE SURROUND COMPUTING ERA

Mark Papermaster
Senior Vice President and CTO, AMD

Hot Chips Symposium
Cupertino, CA
August 28, 2012

AMD

A RAPIDLY CHANGING ENVIRONMENT

Users want content anytime, any platform, anywhere

Explosion of unstructured data

- 245 exabytes of data crossed Internet in 2010[1]
- Growing to 1000 exabytes in 2015

Data center server demand >10M units by 2016[2]



1. Cisco Visual Networking Index Global IP Traffic Forecast, 2010 to 2015
2. Worldwide and Regional Server 2012-2016 Forecast, IDC, May 2012

AMD

REVOLUTIONARY TRANSFORMATION

10 years ago: **The Interactive Computing Revolution**

- Graphics acceleration enabled
- Computing accessible to everyone
- Touch screen phones to cinematic 3D

Starting now: **The Surround Computing Era**

- Computers are everywhere
- Integrating into our environment
- Computing is part of everyday life, not a distinct activity



AMD

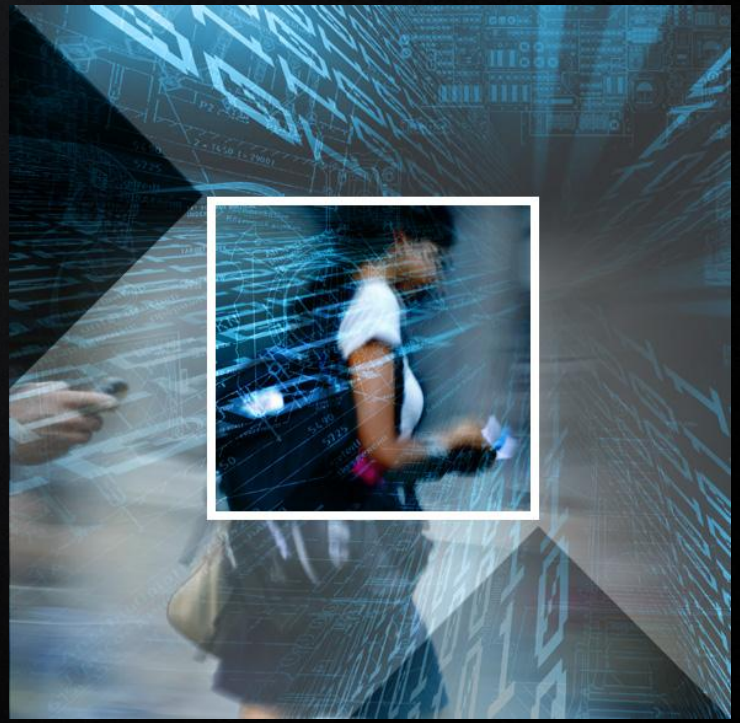
SURROUND COMPUTING

We are entering the Surround Computing Era

- Multi-platform – eyeglasses to room-size
- Fluid – realistic output, natural human input
- Intelligent – anticipates our needs

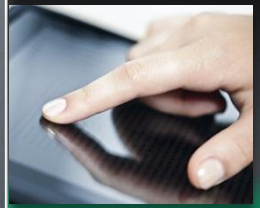
Profound implications for computer architecture

- Smarter clients – realistic, natural human communication
- Smarter clouds – orchestrate 10B devices in real-time



AMD

SMARTER CLIENTS



Natural UI and Gestures

Touch, gesture and voice



Biometric Recognition

Secure, fast, accurate: face, voice, fingerprints



Augmented Reality

Superimpose graphics, audio, and other digital information as a virtual overlay



Content Everywhere

Content from any source to any display seamlessly



Beyond HD Experiences

Streaming media, new codecs, 3D, transcode, audio



AV Content Management

Searching, indexing and tagging of video and audio. Multimedia data mining

New Surround Compute Applications and Experiences – Accelerators Required!

AMD

SMARTER CLOUDS

The Cloud is
the “Backbone” of
Surround Computing

Surround Computing Cloud Services

Trust

Context

Analytic
Compute



Consistent Experiences Across Multiple Devices

AMD

SMARTER CLOUDS – MANAGING UNSTRUCTURED DATA

Connected devices drive cloud computational loads

Datacenter optimized for Surround Computing

- **Scale** – to support tens of billions of connected devices
- **Acceleration** – back-end NUI, graphics, analytics
- **Security, privacy** – consistent end-to-end architecture
- **Real time** – latency is critical
- **Dense servers** – optimized for low power



AMD

THE WAY FORWARD

Surround computing

- Requires smarter clients and clouds
- Efficient datacenters

Heterogeneous engines

- Accelerate key client and server parallel workloads

Heterogeneous System Architecture (HSA)

- New silicon architecture making it all work together



AMD

CHANGING THE THINKING, CHANGING THE GAME

HSA – directly access acceleration hardware

- Unlocks the value of the GPU to software developers
- Program in C, C++, Java, Python, JavaScript, HTML5
- ISA agnostic

GPU = CPU in terms of processing capability

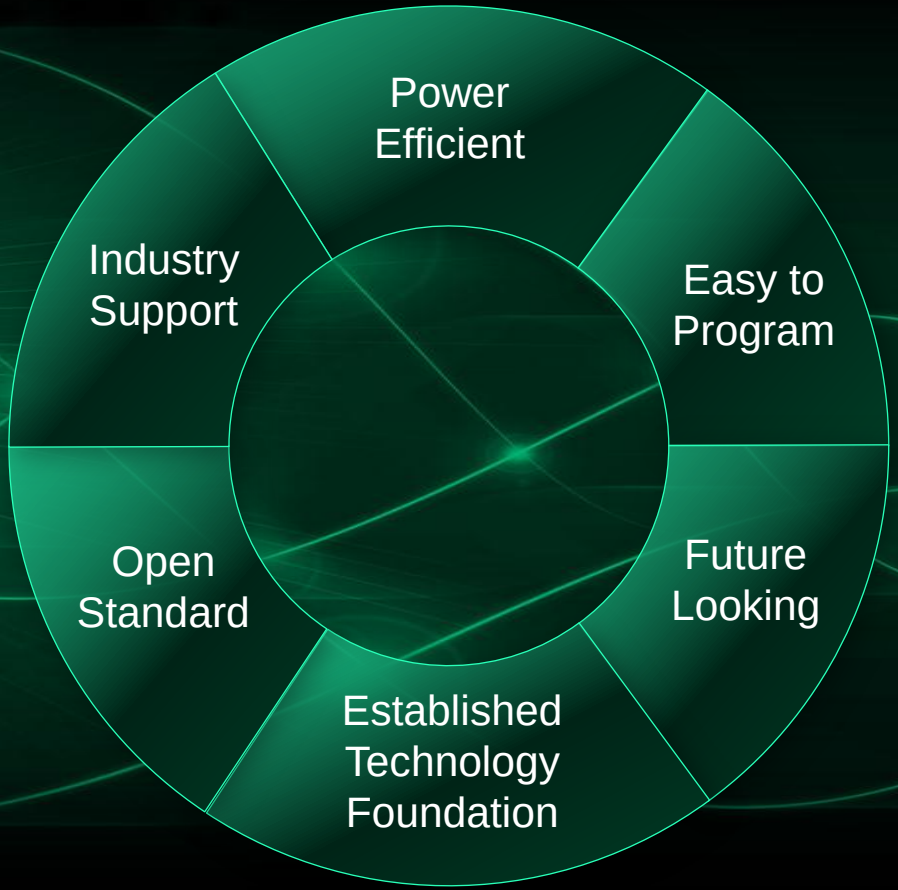
- Full programming language features
- Shared virtual memory: pointer is a pointer
- Coherency and context switching

HSA Foundation is an industry-wide initiative



AMD

BENEFITS OF HETEROGENEOUS SYSTEM ARCHITECTURE



AMD

HSA MEANS ACCELERATED PROCESSING UNITS (APU)

APU is the breakthrough app enabler

APU enables parallel compute and HSA

Emerging workloads require:

- **Seamless execution across CPU/GPU**
- **Other specialized engines**

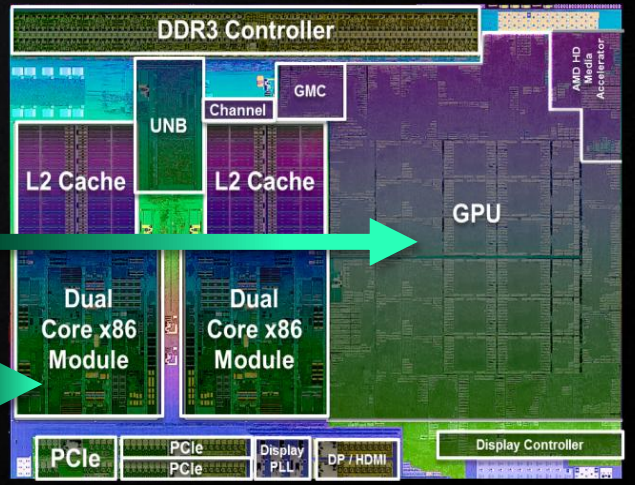
APU is the platform of choice



Data Parallel Workloads

Serial and Task Parallel Workloads

HSA Accelerated Processing Unit



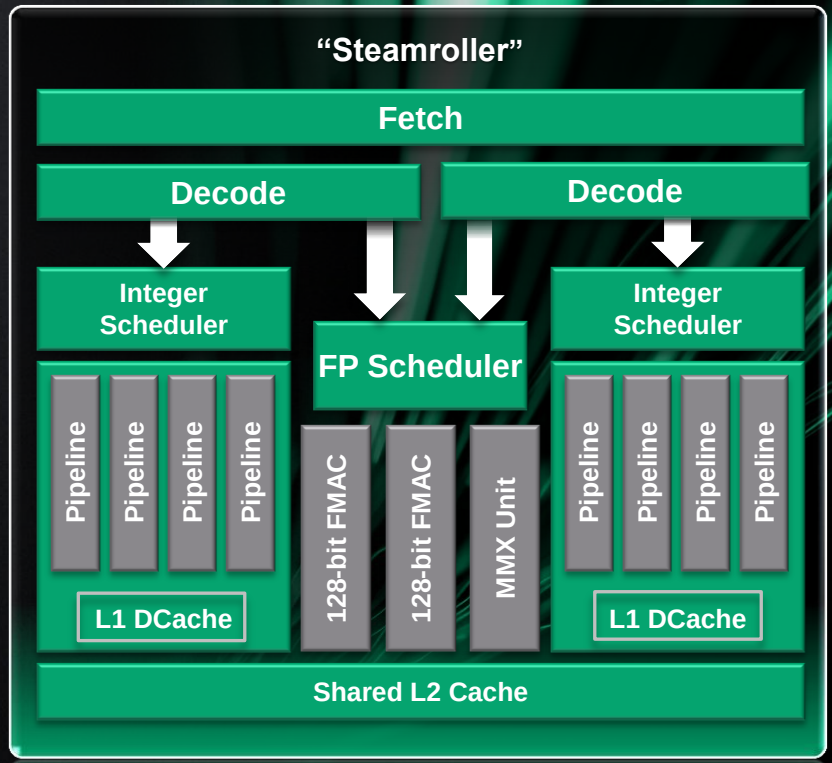
AMD

AMD “STEAMROLLER” CORE

Multi-threaded microarchitecture

Expands computation efficiency

- Feed the cores faster
- Improve single-core execution
- Push on performance/watt

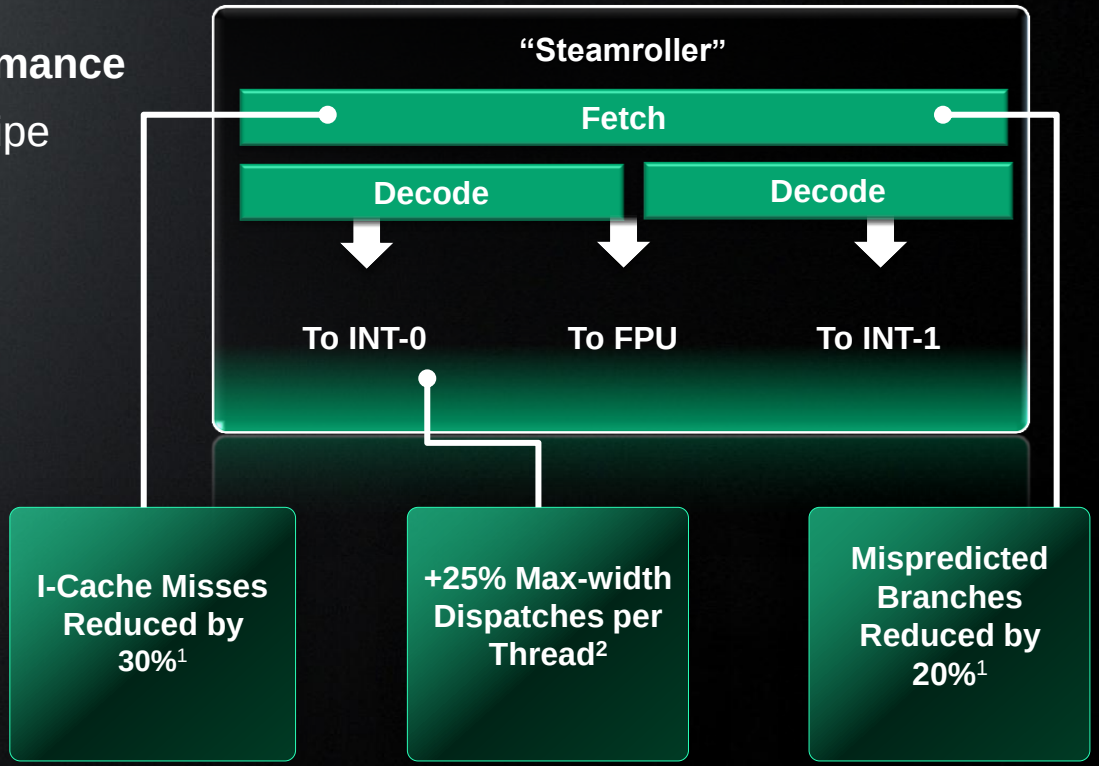


"STEAMROLLER": FEED THE CORES FASTER

No compromises two thread performance

- Dedicated decode for each integer pipe
- Increase instruction cache size
- More efficient dispatch
- Enhance instruction pre-fetch

30% Ops per Cycle Improvement[2]



1.Based on AMD's internal simulation results of average workloads of simulated performance on a number of tests, including those testing transaction processing. (Systems have to be publicly available to publish SPEC CPU Rate.)
2.Based on AMD's internal simulation results of average workloads of simulated performance on a number of tests, including those digital media, productivity and gaming applications.

AMD

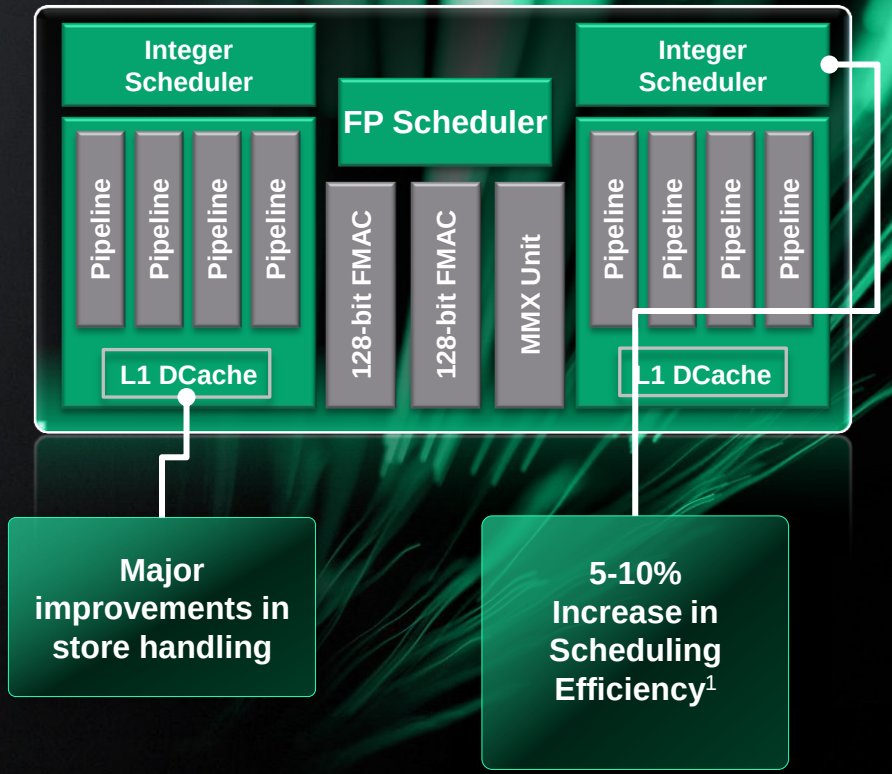
"STEAMROLLER": IMPROVING SINGLE-CORE EXECUTION

Design to tune up integer execution bandwidth:

- **In concert with feeding the core faster**
- More register resources, same latency
- More intelligent scheduling

Design to decrease average load latency:

- **Minimum latency is only part of story**
- Faster handling of data cache misses
- Accelerate store-to-load forwarding



AMD

"STEAMROLLER" PERFORMANCE/WATT DESIGN

Microarchitectural power optimization

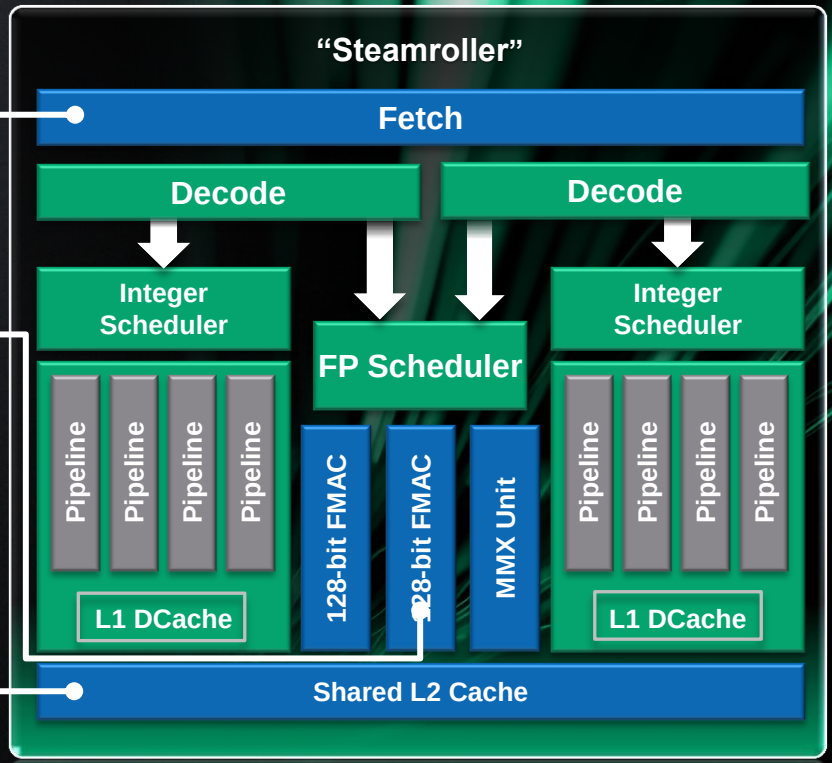
- Lower average dynamic power
- Optimize for loop behaviors

Floating point rebalance

- Streamlined execution hardware
- Adjust to application trends

Dynamic resizing of L2 cache

- Adaptive mode based on workload



AMD

SMART DESIGNING FOR LOW POWER

Power efficiency is fundamental

- Long battery life
- Sleek, light weight form factors
- Cool and quiet computation
- Lower energy consumption and utility bills
- Lower data center TCO

Multi-faceted attack beyond process technology

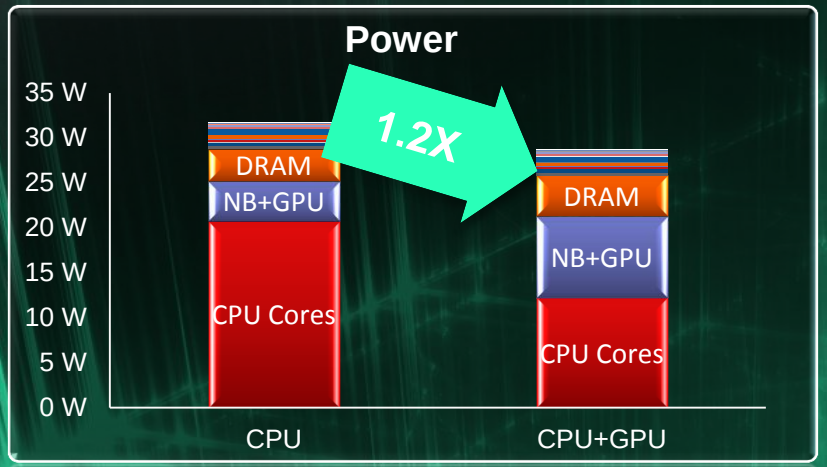
- Optimize hardware with software applications
- Intelligent on-die power management
- Efficient design methodologies

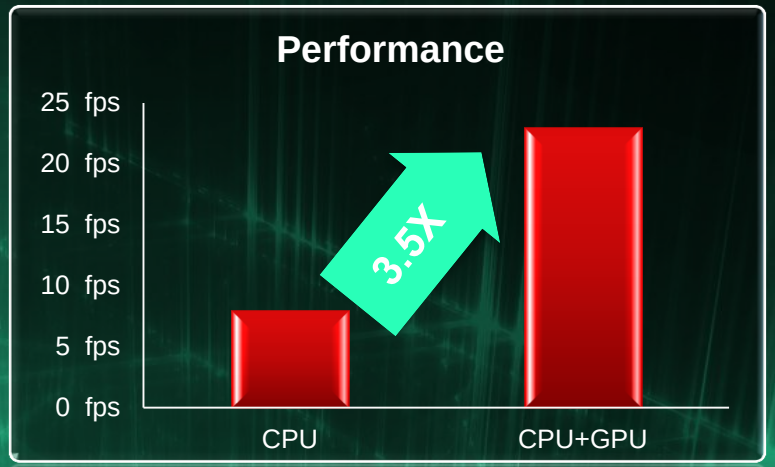


AMD

ARCHITECTURAL EFFICIENCY EXAMPLE WITH VIDEO ENHANCEMENT

MOTION DSP 720P





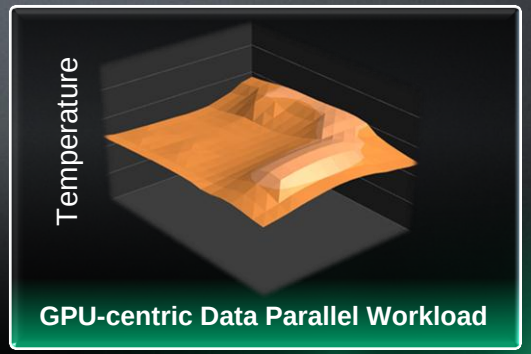
Synergistic use of GPU compute + shared memory = lower power *and* higher performance

>4.0X Better Energy Efficiency[1]

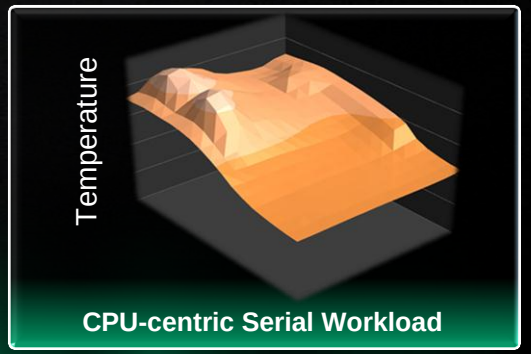
1. AMD ES-3200 APU)Llano-32nm, 2 cores @ 2400Mhz, GPU:2 CU @ 444Mhz), Windows 7 OS, MotionDSP vReveal Applications (http://www.vreveal.com/stabilization)

POWER MANAGEMENT: FINE GRAINED DYNAMIC POWER TRANSFERS

AMD incorporates activity-based power transfer between CPU and GPU

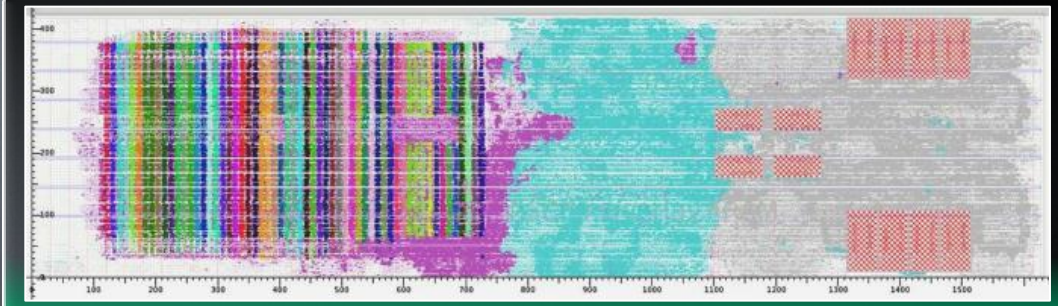






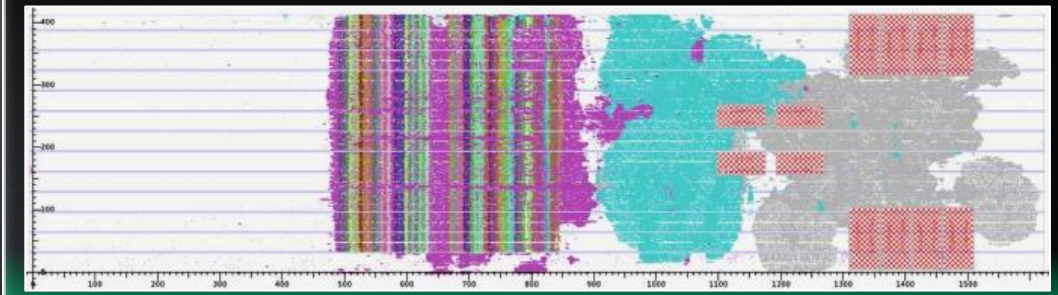
Enabled by sophisticated on-die microcontroller and sensors

POWER EFFICIENCY GAINS FROM IMPROVED DESIGN METHODS



“Bulldozer”

Part of the Floating Point Unit. Hand-drawn for maximum speed and density in 32nm



With High Density Library

The same blocks again, but rebuilt using a **High-Density** cell library to achieve **30% area and power reductions**

15-30% lower energy per operation[1] for power constrained designs – same order as a full process node improvement

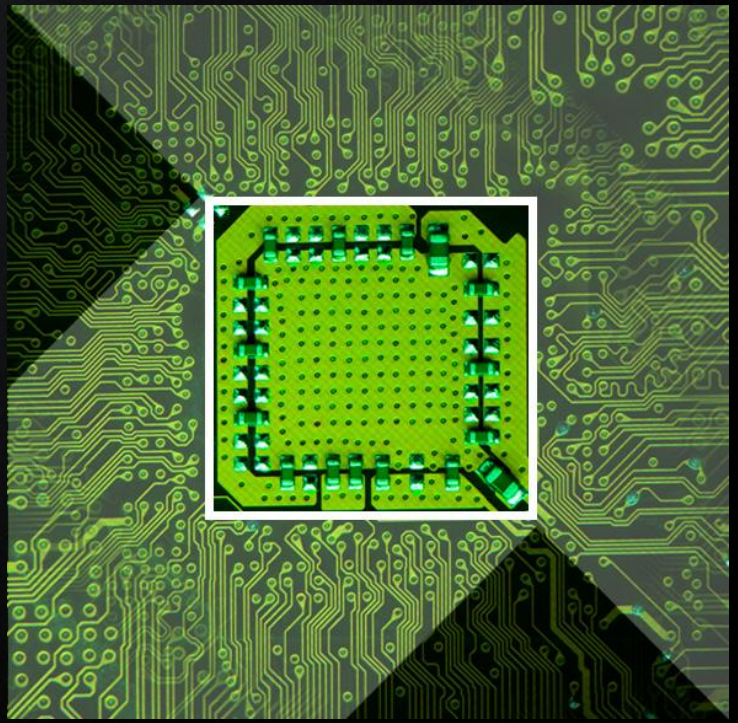
1. Based on AMD’s internal analysis

AMD

FAST FABRICS TIE EVERYTHING TOGETHER

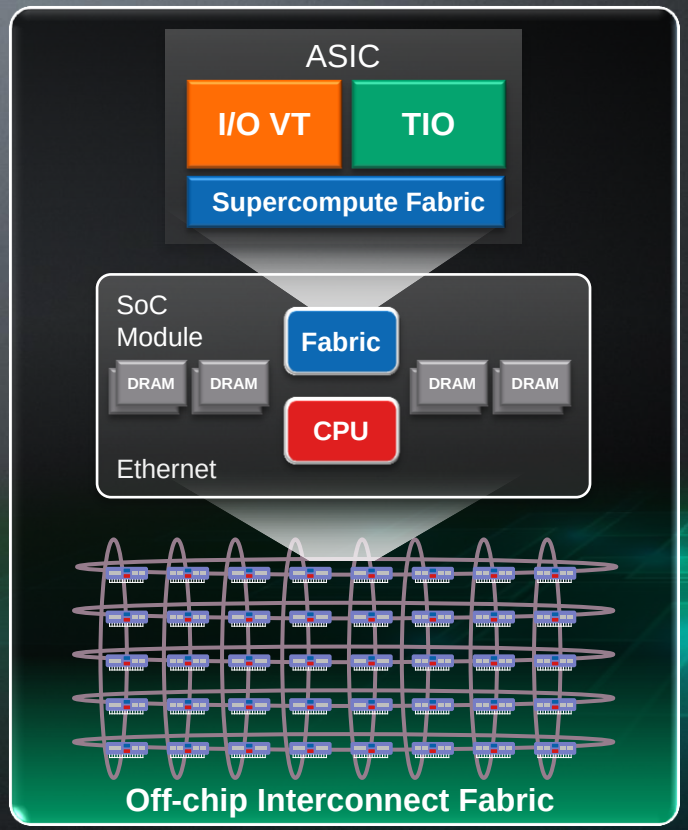
Great interconnect fabrics are needed

- Optimally process unstructured data
- Able to connect massive numbers of processors
- Lowest possible overhead



AMD

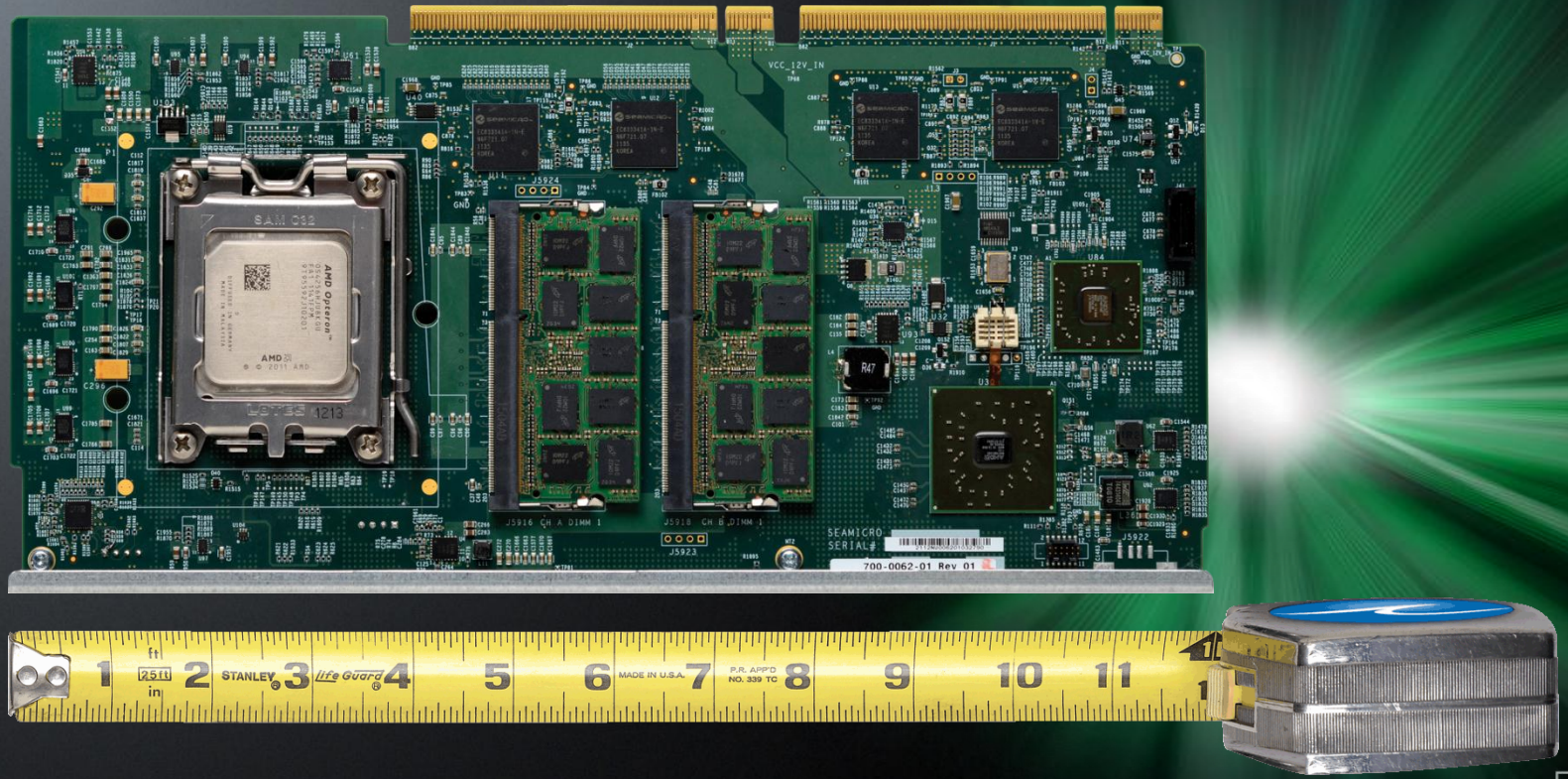
AMD FREEDOM FABRIC™ TECHNOLOGY



AMD off–chip interconnect fabric IP

- Designed to enable significantly lower TCO
- Links hundreds ➡ thousands of SoC modules
- Shares hundreds of TBs storage and virtualizes I/O
- 160Gbps Ethernet Uplink
- Instructions Set Architecture agnostic

END-TO-END SYSTEM OPTIMIZATION



THE PURE SPEEDS AND FEEDS RACE IS OVER – IT'S ABOUT THE SOLUTION!

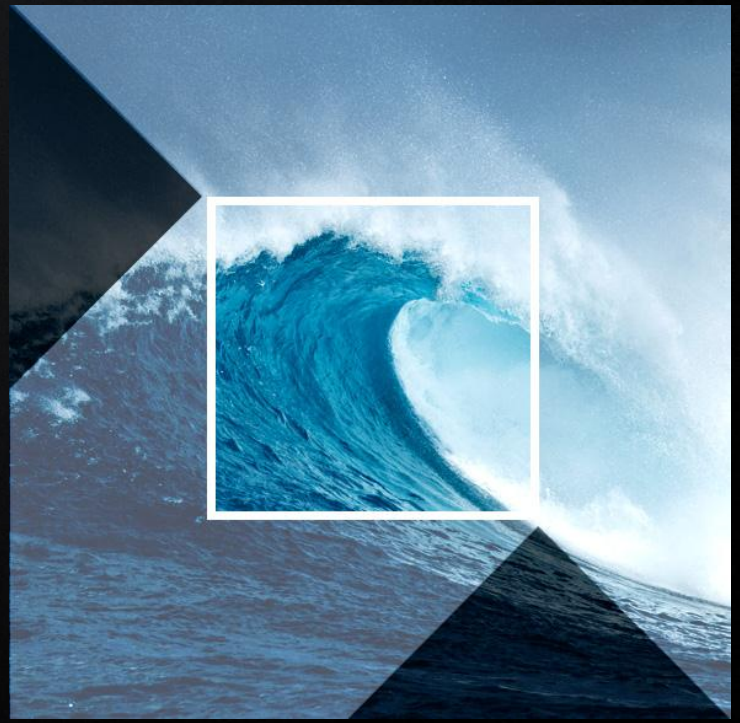
- **End-to-end system view**
- **Acceleration of the application stack**
- **Agile delivery of tailored solutions**
- **Leveraging differentiated IP**



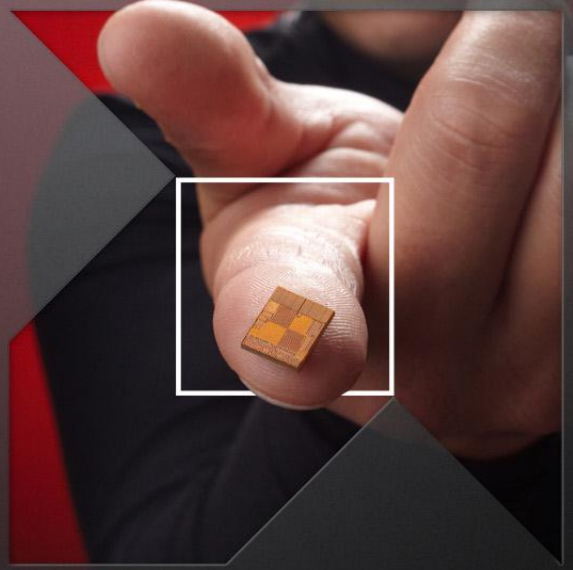
AMD

THE NEXT WAVE – SURROUND COMPUTING REVOLUTION

- AMD products will enable the transition
 - HSA
 - Ambidextrous
 - Fast fabrics
 - Relentless focus on power efficiency
- AMD inspired the interactive computing revolution
- Now leading the way to surround computing



AMD



THANK YOU

AMD

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

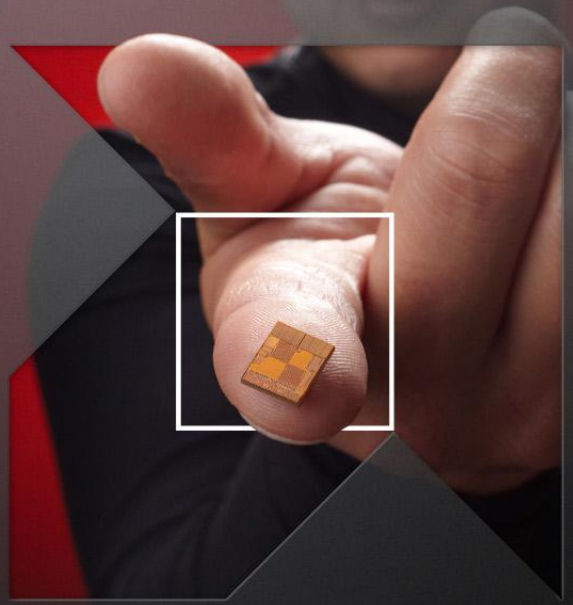
The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

ATTRIBUTION

© 2012 Advanced Micro Devices, Inc. All rights reserved. AMD, the AMD Arrow logo, Radeon, and combinations thereof are trademarks of Advanced Micro Devices, Inc. Other names and logos are used for informational purposes only and may be trademarks of their respective owners.



AMD RADEON™ HD 7970 WITH GRAPHICS CORE NEXT (GCN) ARCHITECTURE

Mike Mantor, AMD Senior Fellow
michael.mantor@amd.com
August 28, 2012

AMD

GRAPHICS CORE NEXT ARCHITECTURE

- Product Goals
 - Time to Market
 - Maximize Performance/Watt
 - Enable first class GPU compute
 - Simplify GPU programming
 - Improve GPU utilization
 - Provide predictable performance
- Parallel Graphics/Compute Architecture
 - New ISA & Compiler
 - Distributed Compute Units
 - Global Unified Read/Write Cache
 - Asynchronous Compute Engines (ACE)
 - Reliability improvements with ECC
- AMD Eyefinity Display Technology
 - Multiple Display Configurations
 - 3D Stereo Displays
 - Flexible Audio



AMD RADEON™ HD 7970 ARCHITECTURE



Graphic Core Next (GCN)

- 4.3 billion 28nm transistors

AMD RADEON™ HD 7970 ARCHITECTURE



Graphic Core Next (GCN)

Advanced Power Management

- Fine grain clock\clock tree gating
- Power Tune – Dynamic V/F Scaling with power containment
- Zero Core Power – Power Gating

AMD RADEON™ HD 7970 ARCHITECTURE

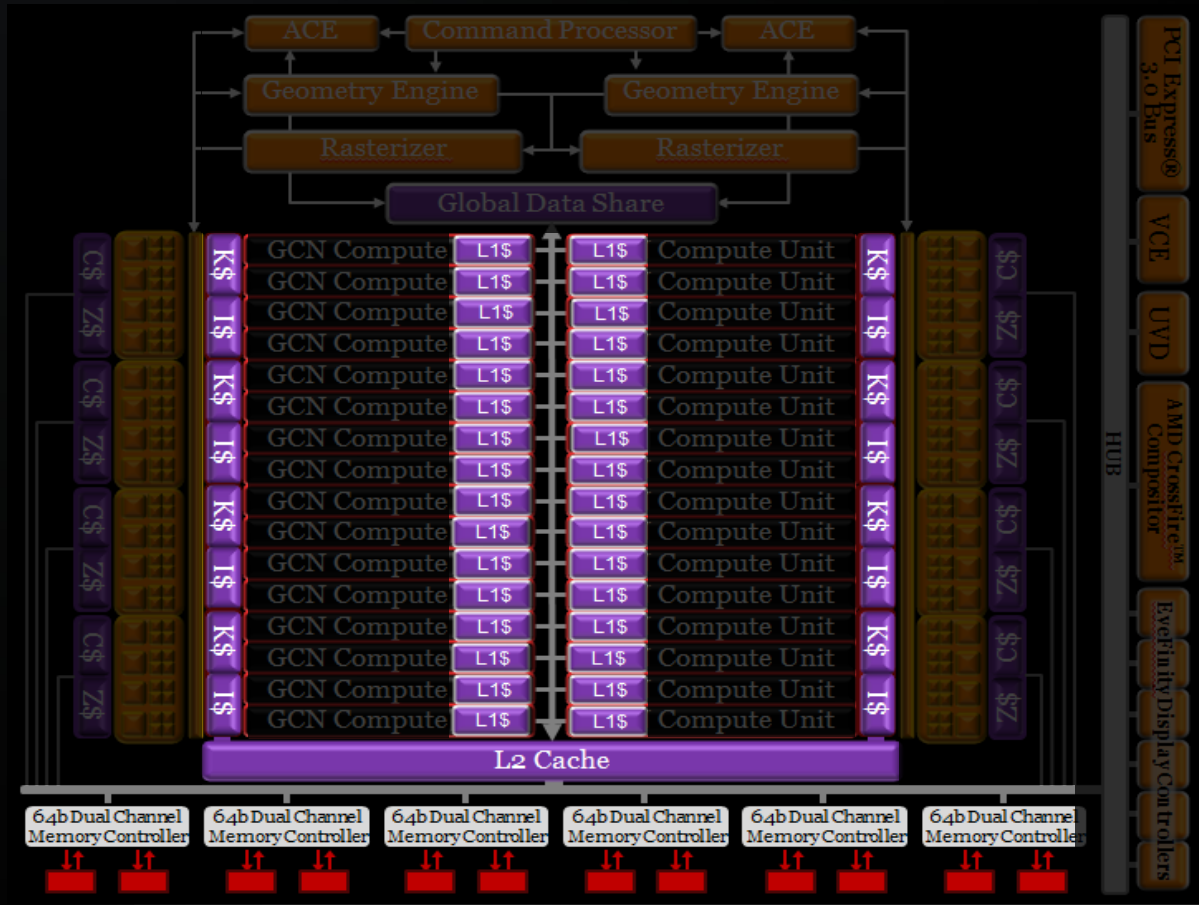


Graphic Core Next (GCN)

32 Compute Units(CU)

- Non VLIW ISA
- Distributed Control Flow
- 32/64b IEEE-2008 FP
- Integer, Logic & Video Ops
- 4 Texture Units per CU

AMD RADEON™ HD 7970 ARCHITECTURE



Graphic Core Next (GCN)

- 384-bit GDDR5 - 264GB/Sec
- Unified R/W Cache Hierarchy
 - 768KB R/W L2 Cache
 - 16KB R/W L1 Per CU
 - 16KB Instruction Cache(I$)/4CU
 - 32KB Scalar Data Cache(K$)/4CU

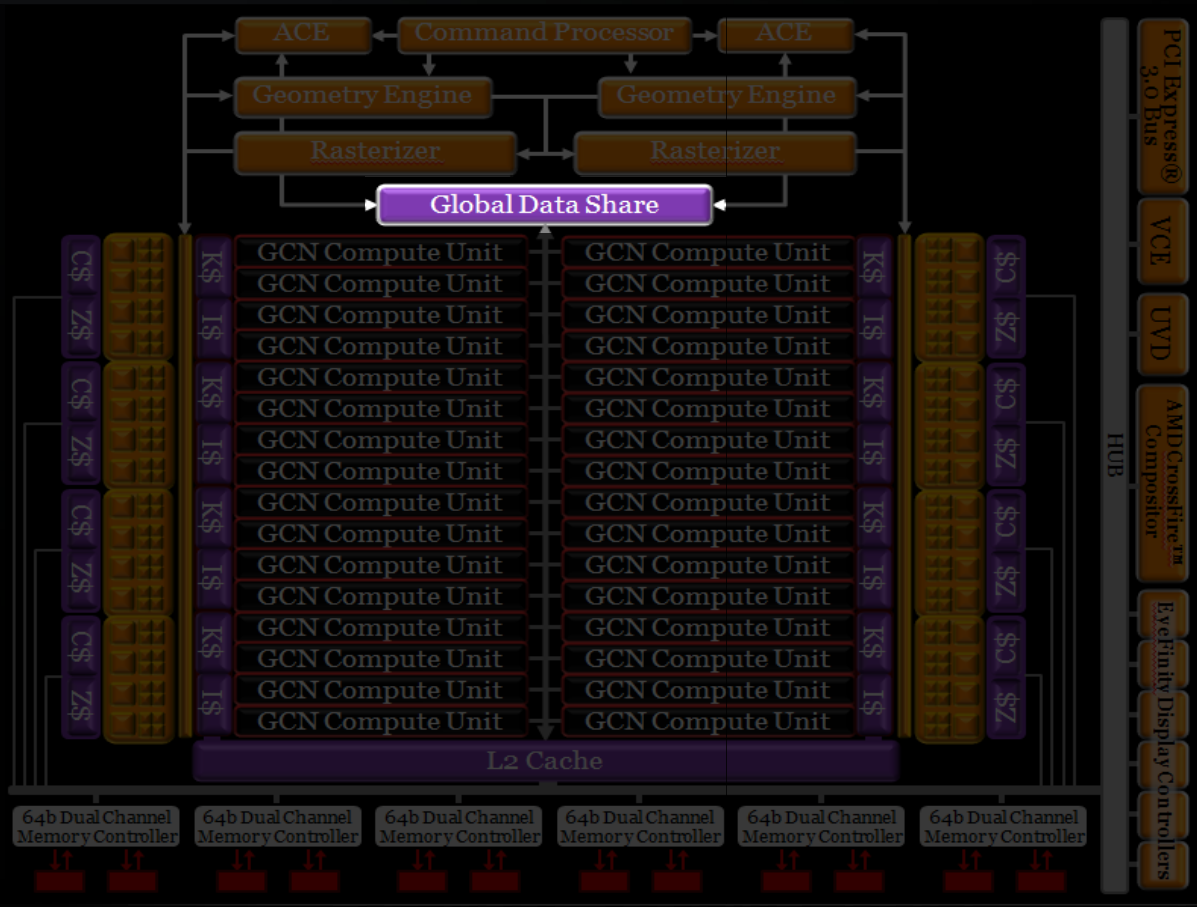
AMD RADEON™ HD 7970 ARCHITECTURE



Graphic Core Next (GCN)

- PCI Express® Gen 3.0 x16

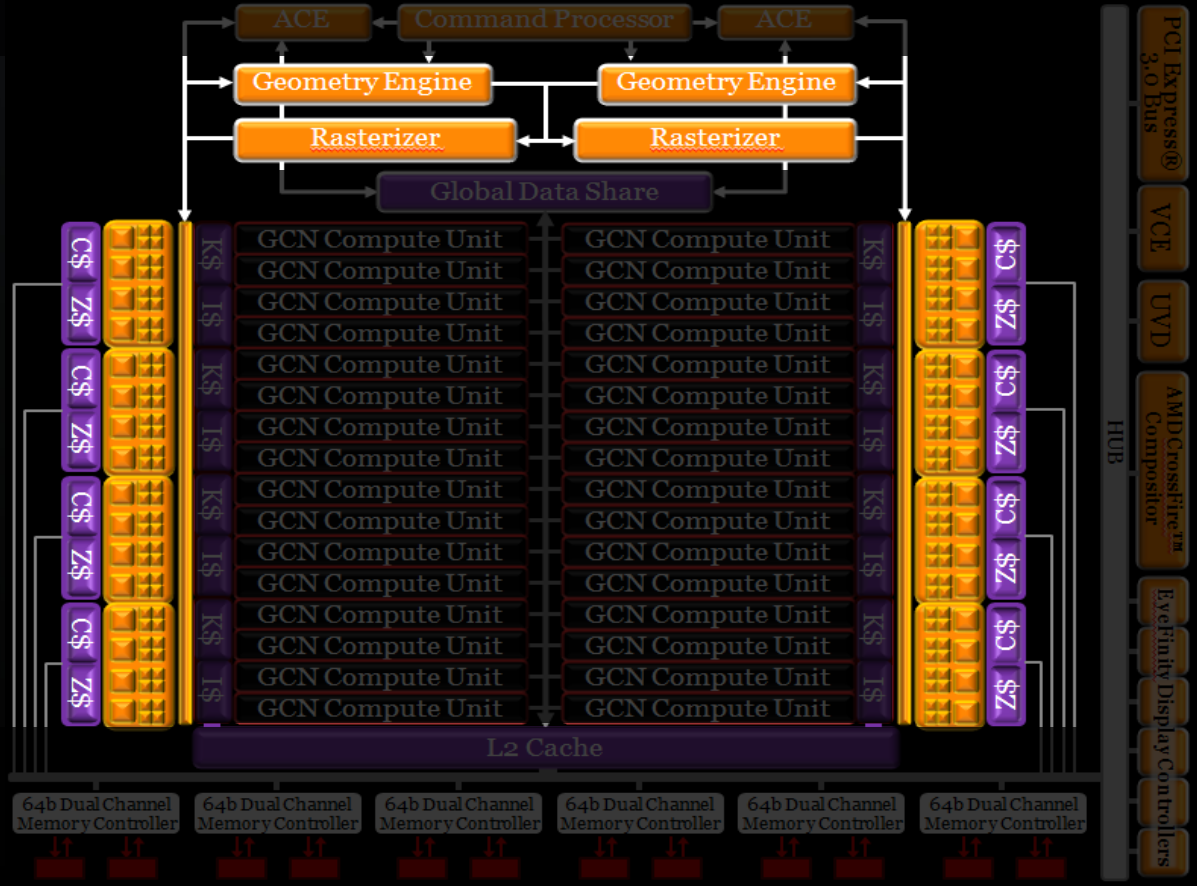
AMD RADEON™ HD 7970 ARCHITECTURE



Graphic Core Next (GCN)

- Global Data Share – 64 kb Shared Memory with global synchronization resources (Barriers, Append, ordered append and named semaphores resources)

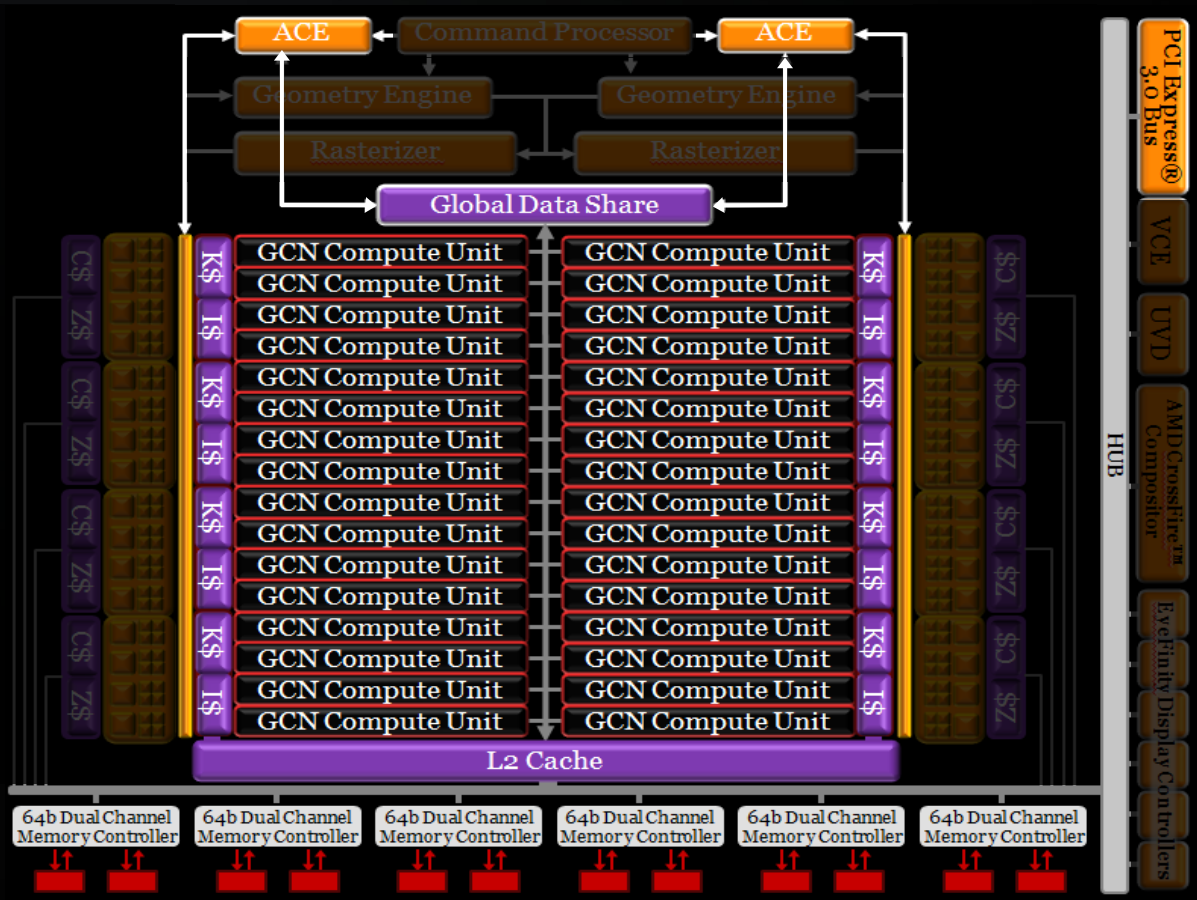
AMD RADEON™ HD 7970 ARCHITECTURE



Graphic Core Next (GCN)

- Dual Geometry Engines
- Dual Rasterizers
- 8 Render Back-ends
 - 32 Pixel Color Raster Operation Pipelines (ROPs)
 - 128 Depth Test (Z)/stencil Ops
 - Color Cache (C$)
 - Depth Cache (Z$)

AMD RADEON™ HD 7970 ARCHITECTURE

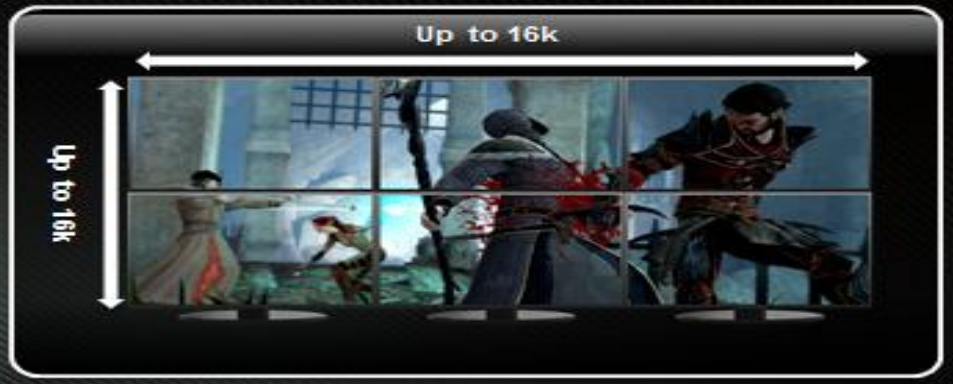


Graphic Core Next (GCN)

- Dual Asynchronous Compute Engines (ACE) and Dual DMA
- Compute ECC protection
 (DRAM & SRAM (Registers, Shared Memories, L1 & L2 Caches)
- GPU support for Compute APIs OpenCL™1.2, DirectCompute, C++ AMP

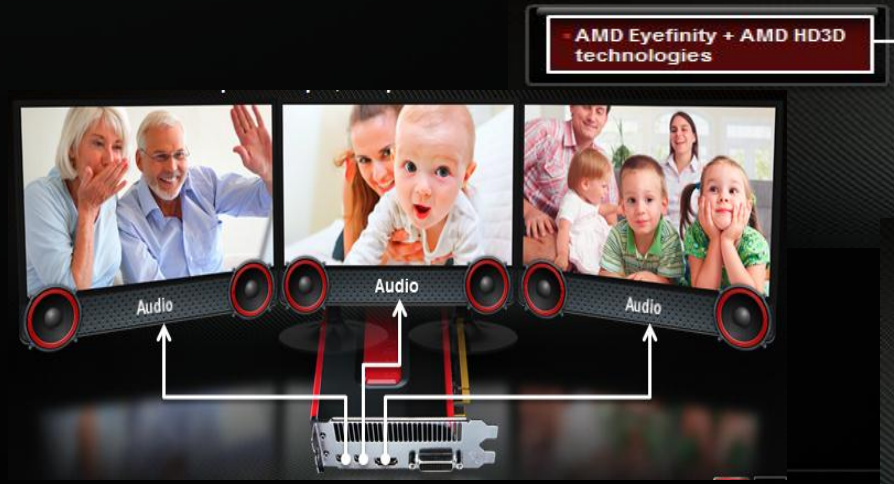
Multi-Media and Display System

- AMD EyeFinity
 - Single 16kx16k Image across 6 Displays
 - Drives three 3D Stereo Display
 - Flexible Bezel Display
- Discrete Digital Multi-Point Audio
- Multi-Display Video Conferencing
- Directional Audio









AMD RADEON™ HD 7970 ARCHITECTURE



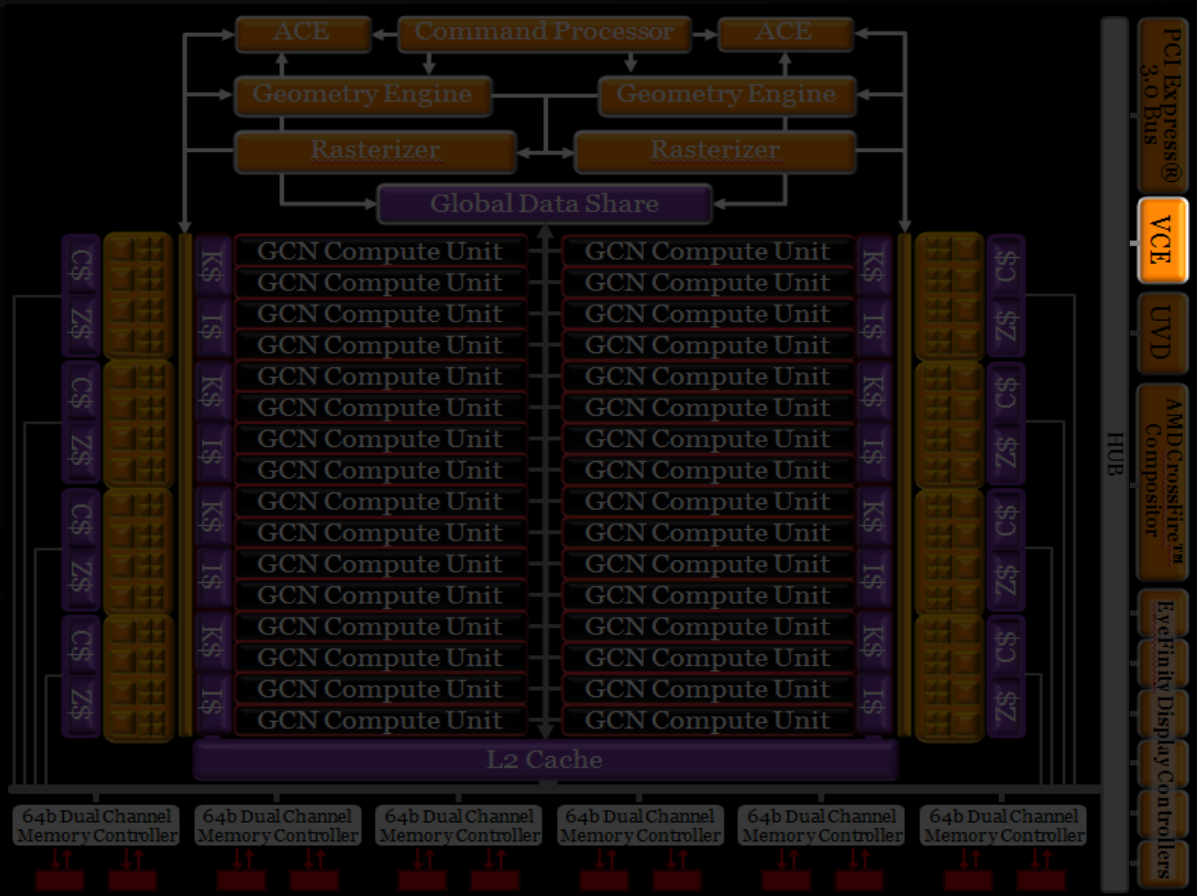
Multi-Media and Display System

Universal Video Decoder (UVD)

Fixed Function with codecs for:

- H.264
- VC-1
- MPEG-2 (SD & HD)
- MVC (Blu-ray HD)
- DivX®
- WMV MFT
- WMV native

AMD RADEON™ HD 7970 ARCHITECTURE

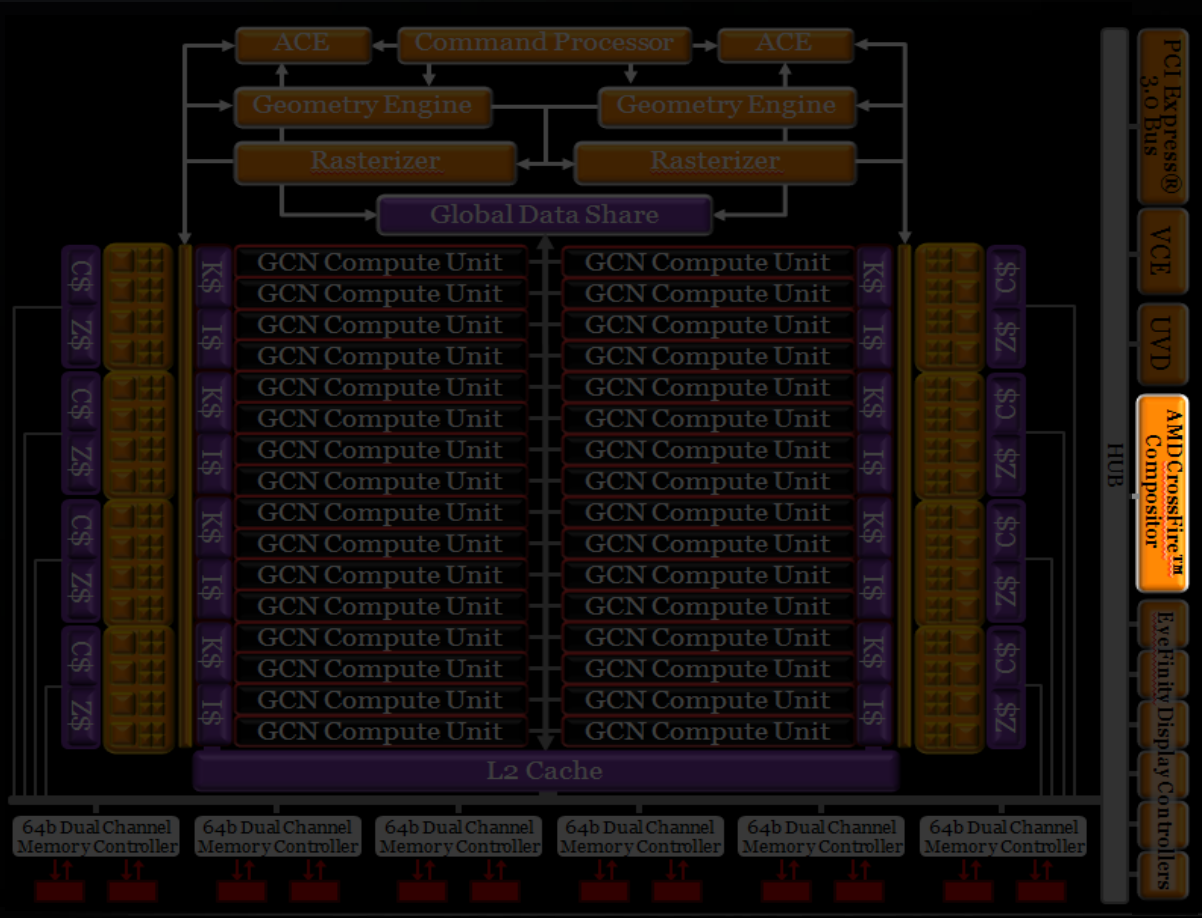


Multi-Media and Display System

Video Codec Engine (Fixed Function)

- Multi-stream hardware H.264 HD Encoder
- Power efficient & faster than real-time 1080p @60fps
- Two encode modes: full fixed & hybrid (with GPU compute)

AMD RADEON™ HD 7970 ARCHITECTURE



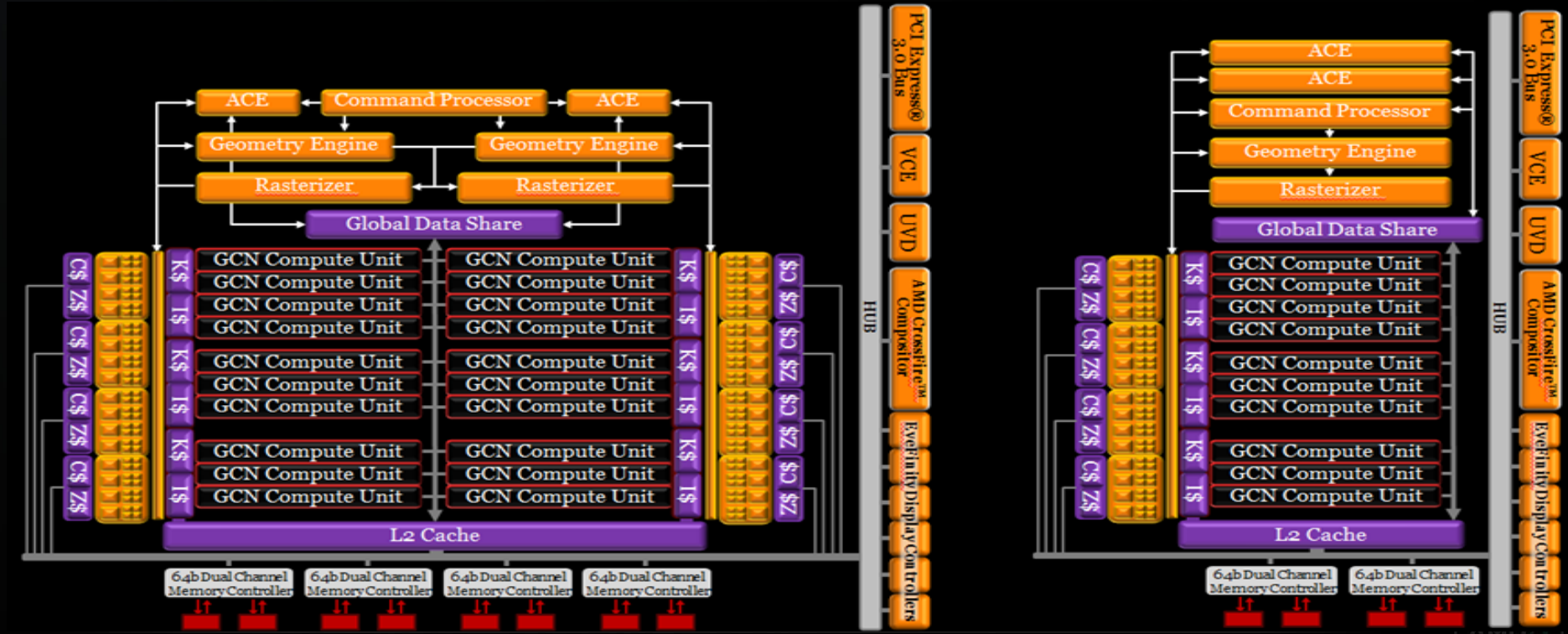
Multi-Media and Display System

CrossFire™ Compositor

- Controller for Multi-GPU Solutions
- Dual, triple or quad-GPU scaling

GCN ARCHITECTURE SUPPORTS MULTIPLE PRODUCT CONFIGURATIONS

- Memory Channels/L2 Partitions & I/O Pins
- Vertex/Primitives/Pixel Rates
- Number of Compute Unit and Number Textures
- 64b Floating Point Rates and ECC Options

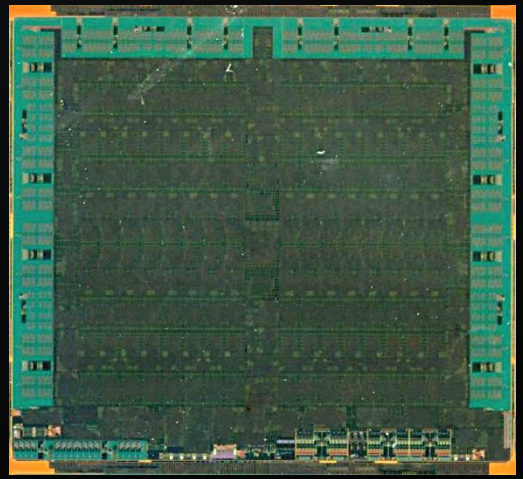




SPEEDS & FEEDS	Dec 2011 AMD Radeon™ HD 7970	Feb 2012 AMD Radeon™ HD 7770	March 2012 AMD Radeon™ HD 7870	June 2012 AMD Radeon™ HD 7970 GHz Edition
Process	28nm	28nm	28nm	28nm
Transistors	4.3 billion	1.5 billion	2.8 billion	4.3 billion
Engine Clock	925 MHz	1000 MHz	1000 MHz	1 GHz / 1.05 GHz
Primitive Rate	2 prim / clk	1 prim / clk	2 prim / clk	2 prim / clk
Stream Processors	2,048	640	1,280	2,048
Compute Performance (SPDP/DPFP)	3.79 TFLOPS / 947 MFLOPS	1.28 TFLOPS / 80 MFLOPS	2.56 TFLOPS / 160 MFLOPS	4.3 TFLOPS / 1.08 TFLOPS
Texture Units	128	40	80	128
Texture Fillrate	118.40 GT/s	40.0 GT/s	80.0 GT/s	134.40 GT/s
ROPS/Pixel Fillrate	32/30.24 GP/s	16/16.0 GP/s	32/32.0 GP/s	32/33.60 GP/s
Z/Stencil	128	64	128	128
Memory Type	3GB GDDR5	2GB GDDR5	2GB GDDR5	3GB GDDR5
Memory Width/Clock	384/1.375 GHz	128/1.125 GHz	256/1.2 GHz	384/1.5 GHz
Memory Data Rate(GDDR5)	5.5 Gbps	4.5 Gbps	4.8 Gbps	6.0 Gbps
Memory Bandwidth	264 GB/s	72 GB/s	153.6 GB/s	288 GB/s
Typical Board Power	~205W	~83W	~150W	~250W
AMD ZeroCore Power	<3W	<3W	<3W	<3W

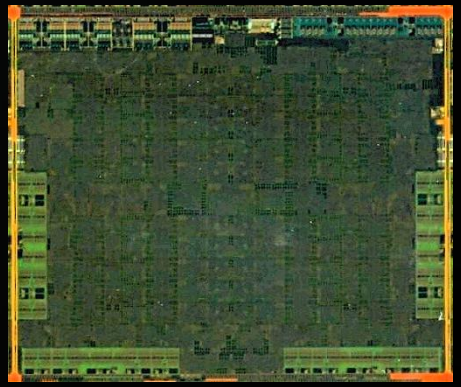
GCN DISCRETE GPU FAMILY

“Tahiti” HD 79XX



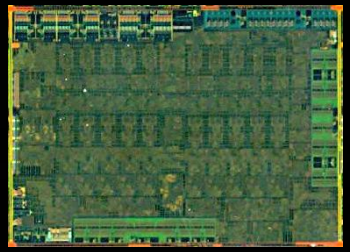
4.3b transistors
352 SQMM
925e Mhz, 3.78 Tflop
32 CU / 32 Pix / 2 Tri
384b , 5.5gbps, 264 GB/S

“Pitcairn” HD 78XX



2.8b transistors
212 SQMM
1e Ghz, 2.56 Tflop
20 CU / 32 Pix / 2 Tri
256b, 4.8gbps, 154 GB/S

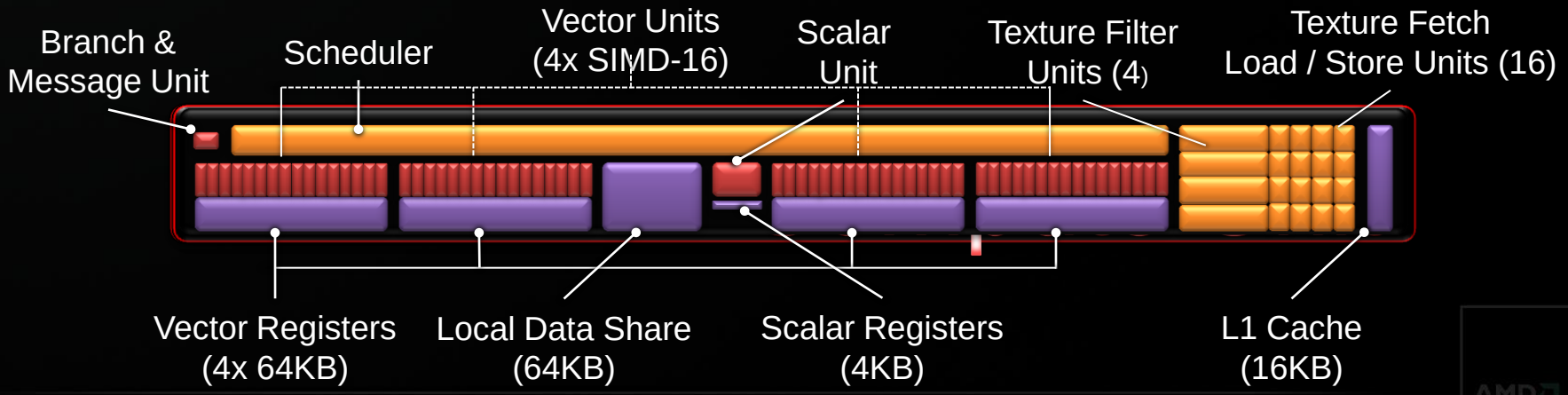
“Verde” HD 77XX



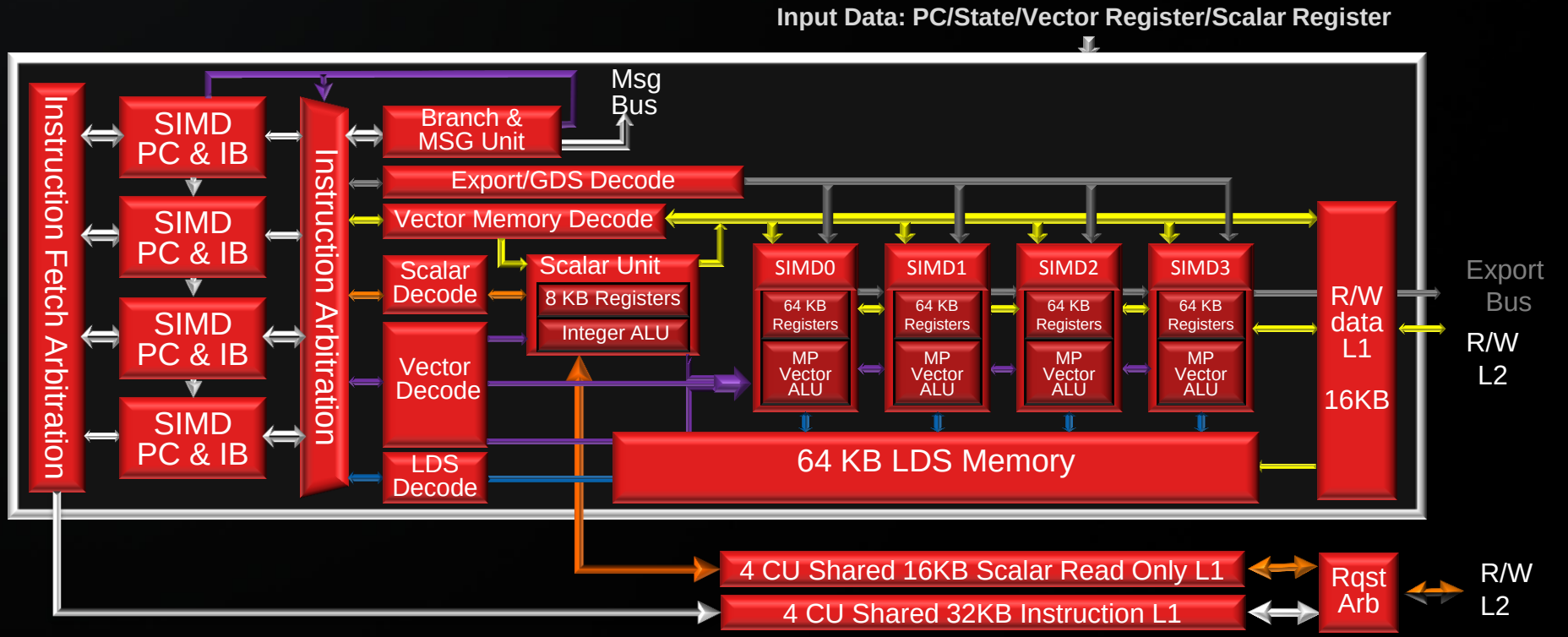
1.5b transistors
123 SQMM
1e Ghz, 1.28 Tflop
10 CU / 16 Pix / 1 Tri
128b, 4.5gbps, 72 GB/S

GCN COMPUTE UNIT

- Basic GPU building block of unified shader system
 - New instruction set architecture
 - Non-VLIW
 - Vector unit + scalar co-processor
 - Distributed programmable scheduler
 - Unstructured flow control, function calls, recursion, Exception Support
 - Un-Typed, Typed, and Image Memory operations
 - Each compute unit can execute instructions from multiple kernels simultaneously
- Designed for programming simplicity, high utilization, high throughput, with multi-tasking

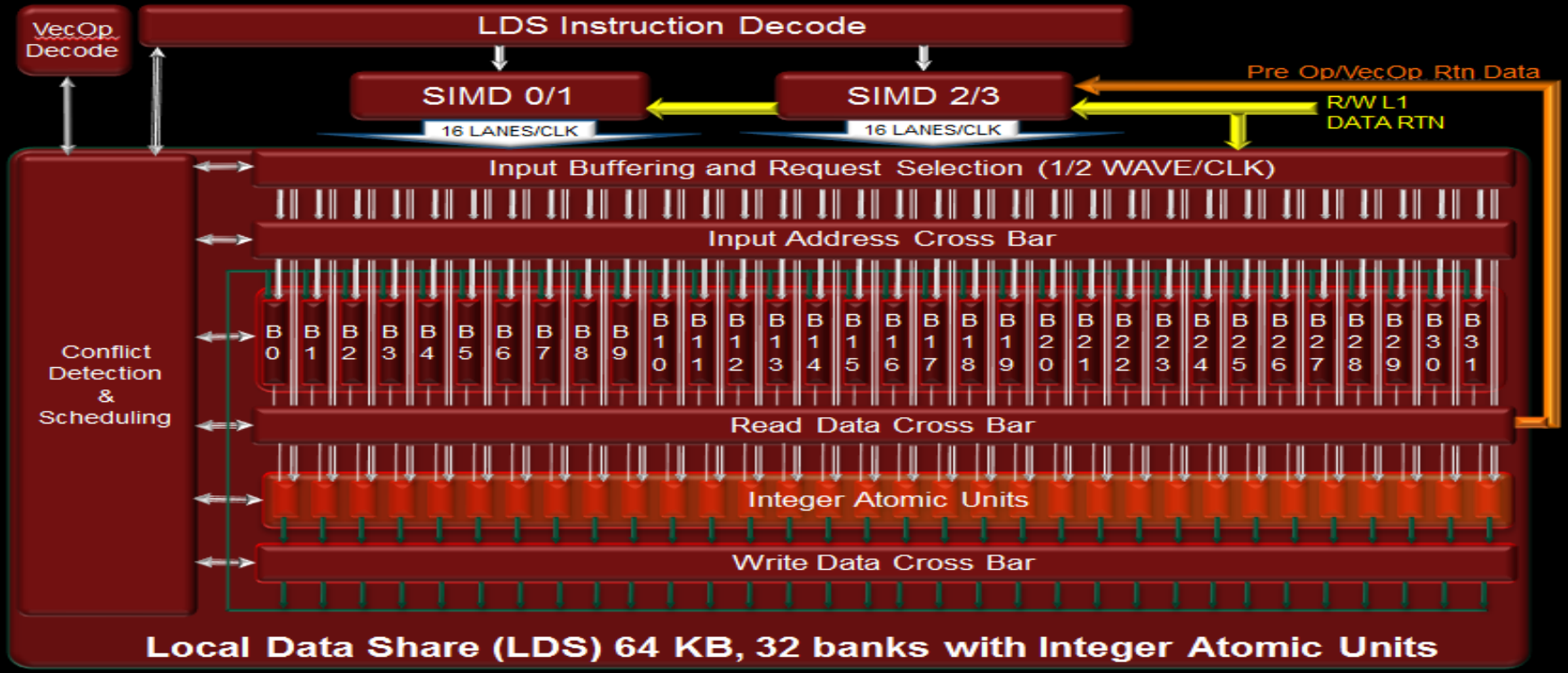


GCN COMPUTE UNIT (CU) ARCHITECTURE



http://developer.amd.com/**afds**/assets/presentations/2620_final.pdf

LOCAL DATA SHARED MEMORY ARCHITECTURE



- 64 kbyte, 32 bank Shared Memory
 - Direct mode – Interpolation @ rate or 1 broadcast read 32/16/8 bit
 - Index Mode – 64 dwords per 2 clks - Service 2 waves per 4 clks
- Advantages
 - Low Latency and Bandwidth amplifier for lower power
 - Software managed cache
 - Software consistency/coherency - thread group via Hardware barrier

PREVIOUS VLIW SHADER ARCHITECTURE

- Previous AMD GPUs used VLIW (Very Long Instruction Word) architecture
 - Combines instructions into a 4-wide VLIW that gets executed on a SIMD

Shader Instructions

```
a = b + c;
b = c + d;
c = d + e;
d = e + f;
```



VLIW Instruction

	X	Y	Z	W	
	b + c	c + d	d + e	e + f	Thread 0
	b + c	c + d	d + e	e + f	Thread 1
	b + c	c + d	d + e	e + f	Thread 2
	⋮	⋮	⋮	⋮	⋮
	b + c	c + d	d + e	e + f	Thread n

PREVIOUS VLIW SHADER ARCHITECTURE

- Previous AMD GPUs used VLIW (Very Long Instruction Word) architecture
 - Combines instructions into a 4-wide VLIW that gets executed on a SIMD

Shader Instructions

With Dependencies

```
a = b + c;
b = a + d;
c = b + e;
d = c + f;
```



VLIW Instruction

	X	Y	Z	W	
	b + c	idle	idle	idle	Thread 0
	b + c	idle	idle	idle	Thread 1
	b + c	idle	idle	idle	Thread 2
	b + c	idle	idle	idle	Thread n

PREVIOUS VLIW SHADER ARCHITECTURE

- Previous AMD GPUs used VLIW (Very Long Instruction Word) architecture
 - Combines instructions into a 4-wide VLIW that gets executed on a SIMD

Shader Instructions

With Dependencies

```
a = b + c;
b = a + d;
c = b + e;
d = c + f;
```



VLIW Instruction

	X	Y	Z	W	
	a + d	idle	idle	idle	Thread 0
	a + d	idle	idle	idle	Thread 1
	a + d	idle	idle	idle	Thread 2
	⋮				⋮
	a + d	idle	idle	idle	Thread n

PREVIOUS VLIW SHADER ARCHITECTURE

- Previous AMD GPUs used VLIW (Very Long Instruction Word) architecture
 - Combines instructions into a 4-wide VLIW that gets executed on a SIMD

Shader Instructions

With Dependencies

```
a = b + c;
b = a + d;
c = b + e;
d = c + f;
```



VLIW Instruction

	X	Y	Z	W	
	b + e	idle	idle	idle	Thread 0
	b + e	idle	idle	idle	Thread 1
	b + e	idle	idle	idle	Thread 2
	b + e	idle	idle	idle	Thread n

PREVIOUS VLIW SHADER ARCHITECTURE

- Previous AMD GPUs used VLIW (Very Long Instruction Word) architecture
 - Combines instructions into a 4-wide VLIW that gets executed on a SIMD

Shader Instructions

With Dependencies

```
a = b + c;
b = a + d;
c = b + e;
d = c + f;
```



VLIW Instruction

	X	Y	Z	W	
	c + f	idle	idle	idle	Thread 0
	c + f	idle	idle	idle	Thread 1
	c + f	idle	idle	idle	Thread 2
	⋮				⋮
	c + f	idle	idle	idle	Thread n

NEW NON-VLIW SHADER ARCHITECTURE

- SIMD architecture without VLIW instructions
 - No need to combine instructions, since multiple threads can run in parallel

Shader Instructions

With or without Dependencies

```
a = b + c;
b = a + d;
c = b + e;
d = c + f;
```

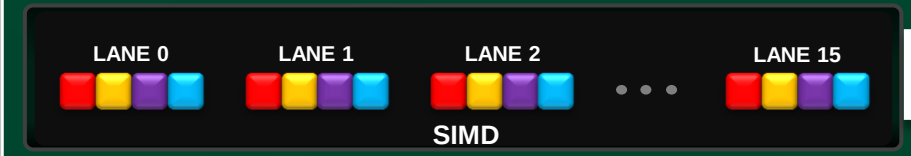


ALUs

S_0	S_1	S_2		S_n
b + c	b + c	b + c		b + c
a + d	a + d	a + d		a + d
b + e	b + e	b + e		b + e
c + f	c + f	c + f		c + f

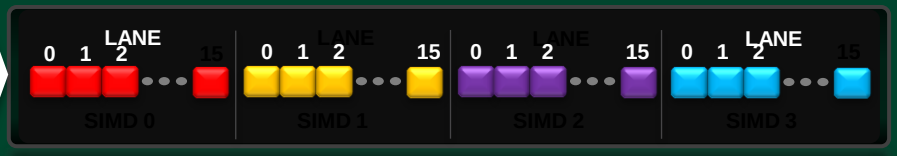
No idle ALUs due to no dependencies!

IS VLIW A GOOD LONG TERM SOLUTION?



VLIW4 SIMD

- 64 Single Precision multiply-add
- 1 VLIW Instruction × 4 ALU ops ➔ dependency limited
- Compiler manages register port conflicts
- Specialized, complex compiler scheduling
- Difficult assembly creation, analysis, and debug
- Complicated tool chain support
- Careful optimization required for peak performance



GCN Quad SIMD

- 64 Single Precision multiply-add
- 4 SIMDs × 1 ALU op ➔ occupancy limited
- No register port conflicts
- Standardized compiler scheduling & optimizations
- Simplified assembly creation, analysis, and debug
- Simplified tool chain development and support
- Stable and predictable performance

VLIW packing sometimes requires domain transformation to achieve good utilization.

CODE EXAMPLE

```
float fn0(float a,float b)
{
  if(a>b)
    return((a-b)*a);
  else
    return((b-a)*b
```

```
//Registers r0 contains "a", r1 contains "b"
//Value is returned in r2

  v_cmp_gt_f32     r0,r1          //a > b, establish VCC
  s_mov_b64        s0,exec        //Save current exec mask
  s_and_b64        exec,vcc,exec  //Do "if"
  v_sub_f32        r2,r0,r1       //result = a - b
  v_mul_f32        r2,r2,r0       //result=result
  s_andn2_b64      exec,s0,exec   //Do "else"(s0 & !exec)
  v_sub_f32        r2,r1,r0       //result = b - a
  v_mul_f32        r2,r2,r1       //result = result * b
  s_mov_b64        exec,s0        //Restore exec mask
```

- Generally straight forward to generate and understand ISA
- VCC - Vector condition code
- EXEC – Execution mask
- Multi-threaded enables full vector unit utilization

GCN SCALAR/VECTOR COMPUTE UNIT

- Simpler ISA compared to previous generation
 - No clauses and latency for transitions
 - No VLIW packing required
 - Control flow directly programmed (Exec mask control)
 - Complex Control Flow Supported (Example: non uniform Branch into loop)
- Scalar engine
 - Lower latency for distributed sequencer verses previous centralized
 - Reduces performance in previously clause bound cases
 - Reduces power handling of control flow Ops as control is closer
- Advanced language feature support
 - Exception support
 - Function calls
 - Recursion
- Enhanced extended ALU operations
 - Media ops
 - Integer ops
 - Integer atomic operations
 - Floating point atomics (min, max, cmpxchg)
- Enhanced debug support
 - HW functionality to improve debug support

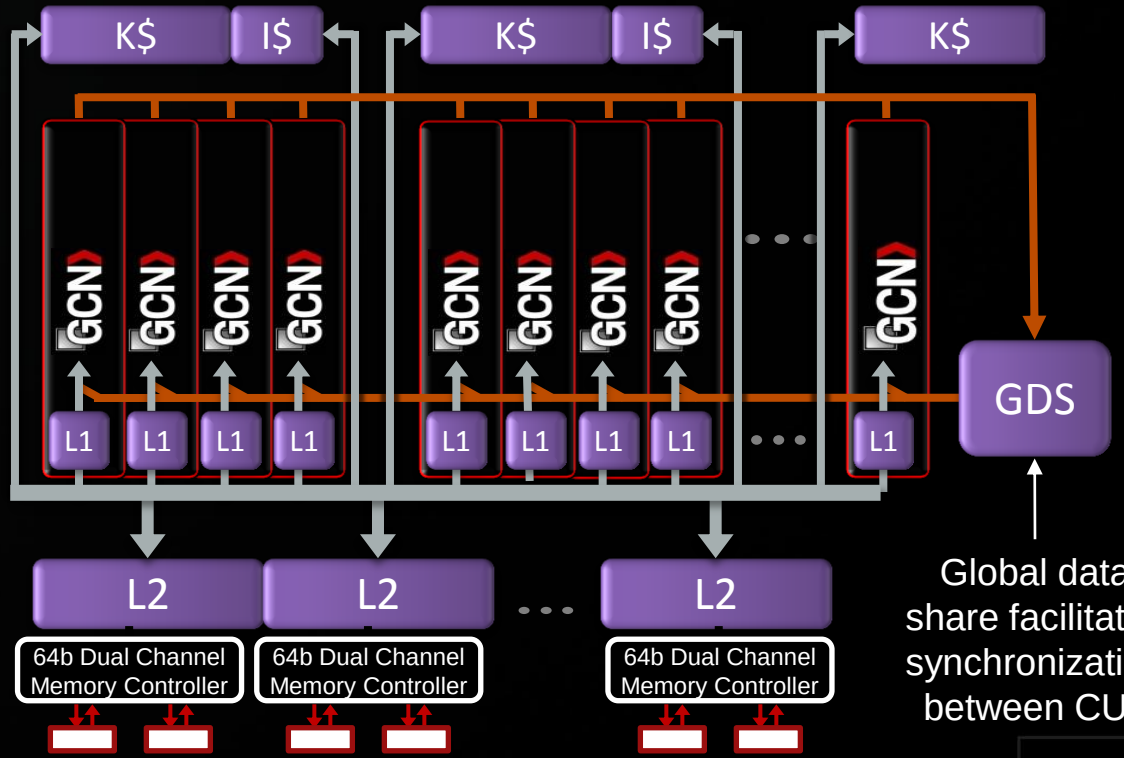
R/W CACHE HIERARCHY

16KB instruction cache (I$) +
32 KB scalar data cache (K$)
shared per 4 CUs with L2 backing

Each CU has 256kb registers and
64kb local data share

L1 read/write 16kb write
through caches
64 Bytes / CU / clock

L2 read/write cache partitions
(64kb/128kb) write back caches
64 Bytes / partition / clock



Global data
share facilitates
synchronization
between CUs

GPU MEMORY MODEL

- Relaxed memory model
 - All work-items within same work groups see same L1 cache
 - Work-items of different work groups may use different L1 caches
 - All work-items and command streams use the same L2 cache
 - Command stream packets & Shader Instruction control data visibility
 - Sufficient primitives in the GPU hardware to implement C++ 11 memory model
- GPU Coherency
 - Acquire/Release semantics control data visibility across the machine (Compiler controlled bit on load/store instructions)
 - L2 coherent ➔ all CUs & CP can have the same view of memory
- Remote Global atomics
 - Performed in L2 cache
 - Full set of integer ops and float max, min, cmp_swap

AMD GCN CU ARCHITECTURE SUMMARY

- Heavily multi-threaded CU architected for throughput
 - Efficiently balanced for graphics and general compute
 - Simplified coding for performance, debug and analysis
 - Simplified machine view for tool chain development
 - Low latency flexible control flow operations
 - Read/Write Cache Hierarchy improves I/O characteristics
 - Flexible vector load, store, and remote atomic operations
 - Load acquire/store release consistency controls



REFERENCE

http://www.amd.com/us/products/desktop/graphics/7000/7970/Pages/radeon-7970.aspx#/1

- AMD Display Technologies whitepaper
- AMD Eyefinity Technology whitepaper
- AMD Power Technologies whitepaper
- AMD Video Technologies whitepaper
- Graphics Core Next Architecture whitepaper

*http://developer.amd.com/**afds**/assets/presentations/2620_final.pdf*

CODE EXAMPLE

```
float fn0(float a,float b)
{
   if(a>b)
      return((a-b)*a);
   else
      return((b-a)*b
```

Optional:

Use based on the number of instruction in conditional section.

- Executed in branch unit

```
//Registers r0 contains "a", r1 contains "b"
//Value is returned in r2

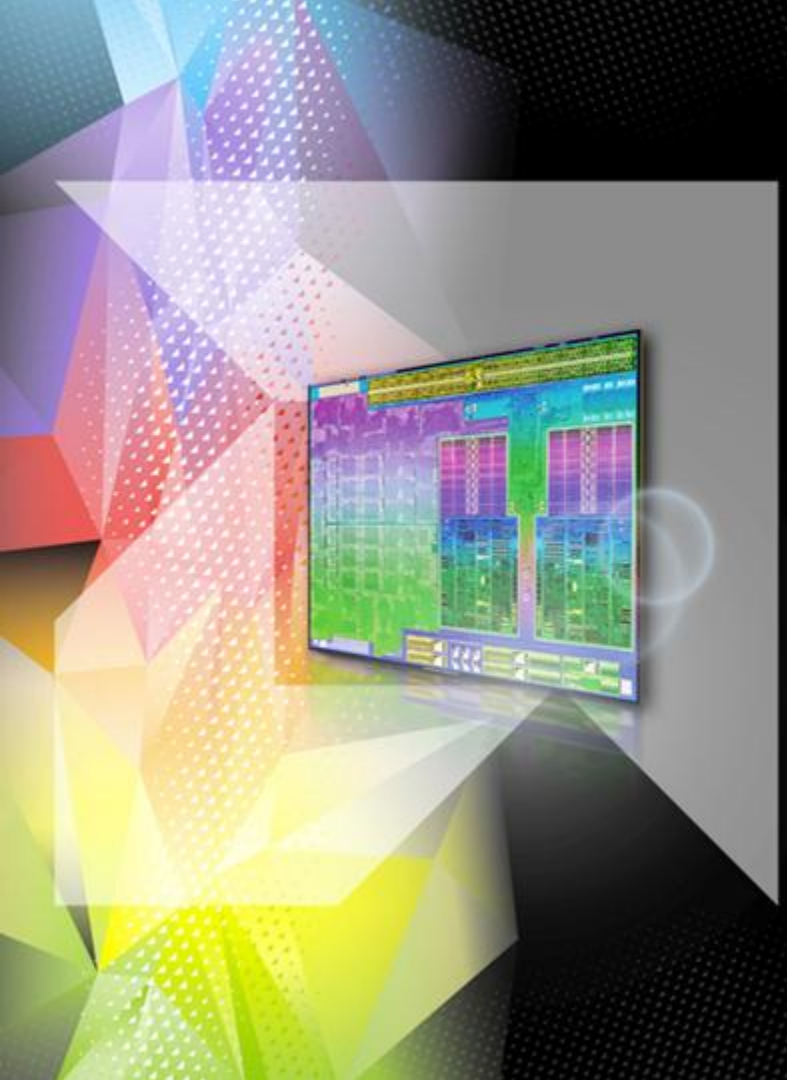
   v_cmp_gt_f32     r0,r1          //a > b, establish VCC
   s_mov_b64        s0,exec        //Save current exec mask
   s_and_b64        exec,vcc,exec  //Do "if"
   s_cbranch_vccz   label0         //Branch if all lanes fail
   v_sub_f32        r2,r0,r1       //result = a - b
   v_mul_f32        r2,r2,r0       //result=result * a
label0:

   s_andn2_b64      exec,s0,exec   //Do "else"(s0 & !exec)
   s_cbranch_execz  label1         //Branch if all lanes fail
   v_sub_f32        r2,r1,r0       //result = b - a
   v_mul_f32        r2,r2,r1       //result = result * b
label1:
   s_mov_b64        exec,s0        //Restore exec mask
```

- Generally straight forward to generate and understand ISA
- Instructions types interleave within program
- Throughput optimized for vector instructions
- Optional scalar instructions jump fully predicated groups of instructions

Disclaimer & Attribution

- The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.
- The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. There is no obligation to update or otherwise correct or revise this information. However, we reserve the right to revise this information and to make changes from time to time to the content hereof without obligation to notify any person of such revisions or changes.
- NO REPRESENTATIONS OR WARRANTIES ARE MADE WITH RESPECT TO THE CONTENTS HEREOF AND NO RESPONSIBILITY IS ASSUMED FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.
- ALL IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE ARE EXPRESSLY DISCLAIMED. IN NO EVENT WILL ANY LIABILITY TO ANY PERSON BE INCURRED FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.
- AMD, the AMD arrow logo, AMD Radeon and combinations thereof are trademarks of Advanced Micro Devices, Inc. All other names used in this presentation are for informational purposes only and may be trademarks of their respective owners.
- OpenCL is a trademark of Apple Inc. used with permission by Khronos.
- DirectX is a registered trademark of Microsoft Corporation.
- DivX is a registered trade mark of DivX Inc
- PCI Express is a registered trademark of PCI-SIG
- © 2012 Advanced Micro Devices, Inc. All rights reserved.



AMD

HOT CHIPS 2012

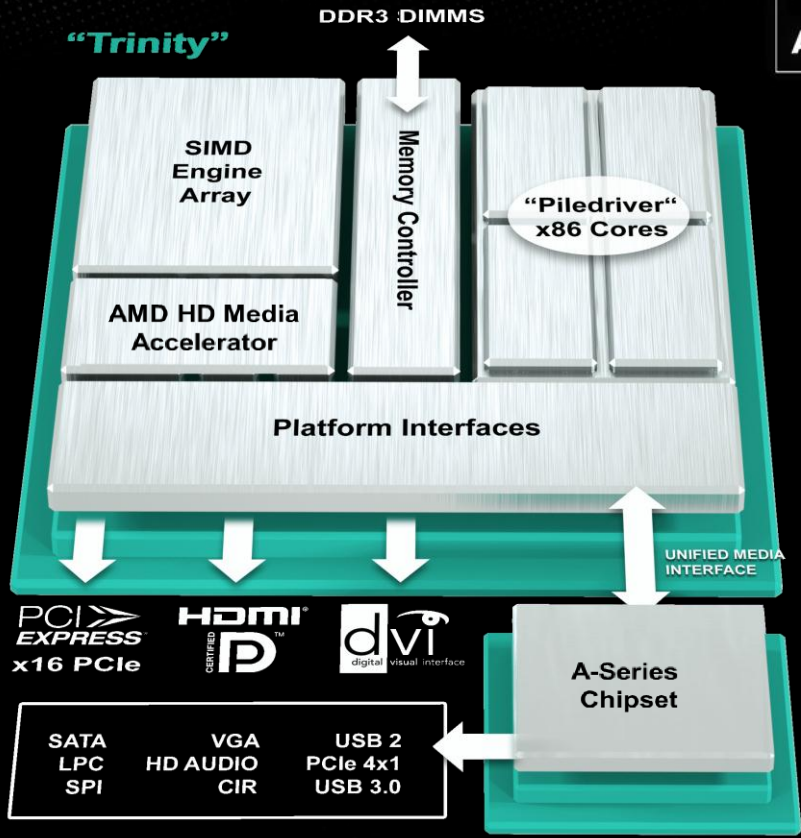
AMD "TRINITY" APU

Sebastien Nussbaum
AMD Fellow
Trinity SOC Architect

AMD APU "TRINITY" WITH AMD DISCRETE CLASS GRAPHICS

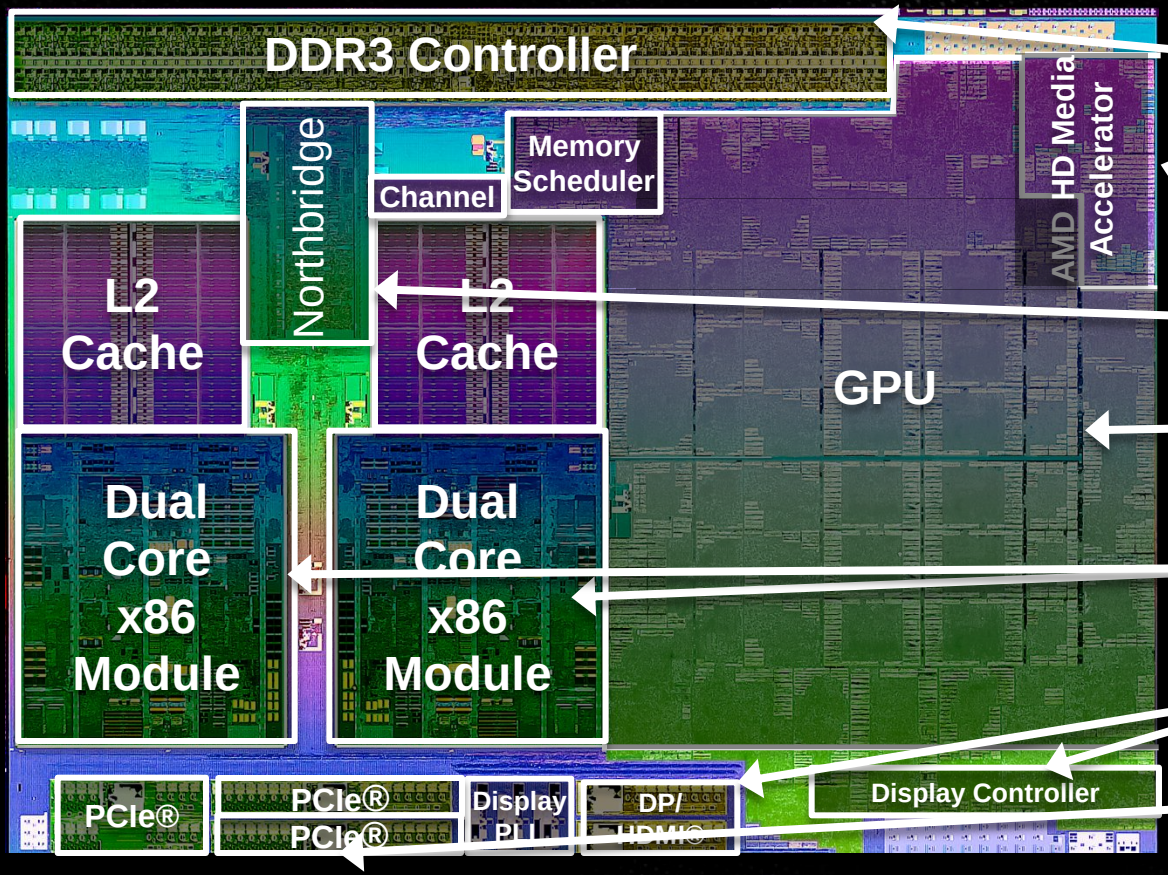
ALL NEW ARCHITECTURE FOR UP TO 50% GPU[1] AND UP TO 25% BETTER X86 PERFORMANCE[2]

- **"Piledriver" Cores**
 - Improved performance and power efficiency
 - 3rd-Gen Turbo Core technology
 - Quad CPU Core with total of 4MB L2
- **2nd-Gen AMD Radeon™ with DirectX® 11 support**
 - 384 Radeon™ Cores 2.0
- **HD Media Accelerator**
 - Accelerates and improves HD playback
 - Accelerates media conversion
 - Improves streaming media
 - Allows for smooth wireless video
- **Enhanced Display Support**
 - AMD Eyefinity Technology[3]
 - 3 Simultaneous DisplayPort 1.2 or HDMI/DVI links
 - Up to 4 display heads with display multi-streaming



"TRINITY" FLOORPLAN

32nm SOI, 246mm2, 1.303BN TRANSISTORS



Dual Channel DDR3 Memory Controller

AMD HD Media Accelerator (UVD, AMD Accelerated Video Converter)

Unified Northbridge

AMD Radeon™ GPU

Up to 4 "Piledriver" Cores with total 4MB L2

HDMI, DisplayPort 1.2, DVI controllers

PCI Express® I/O — 24 lanes, optional digital display interfaces

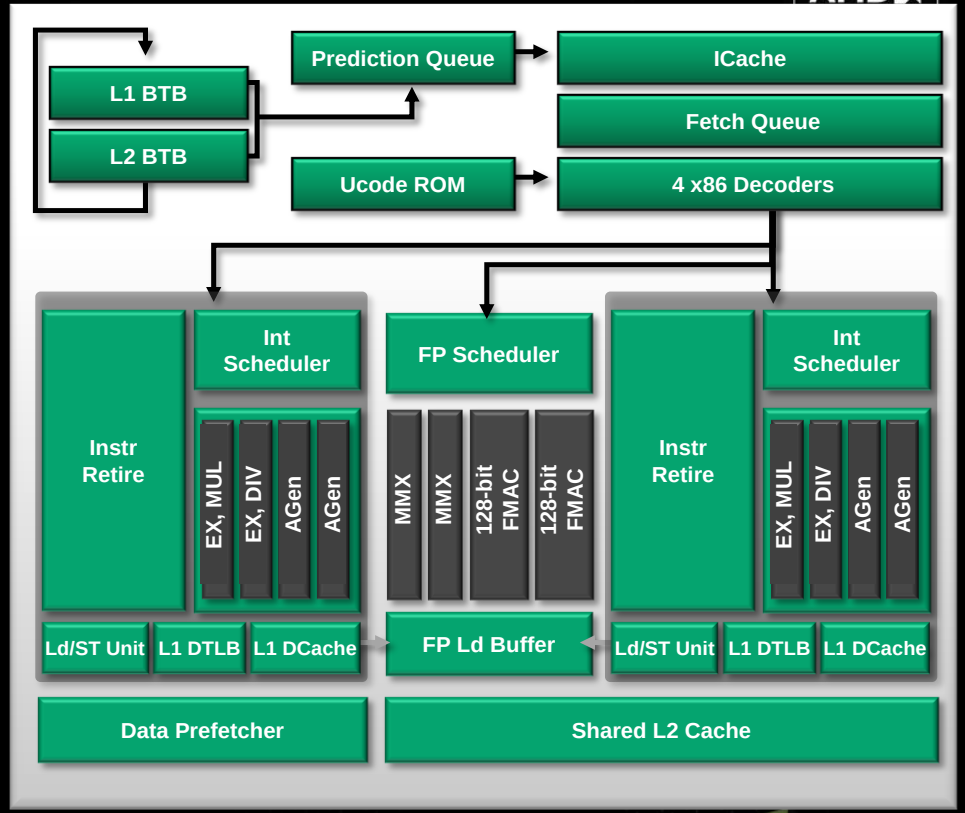
AMD 2ND GENERATION "BULLDOZER" CORE: "PILEDRIVER"



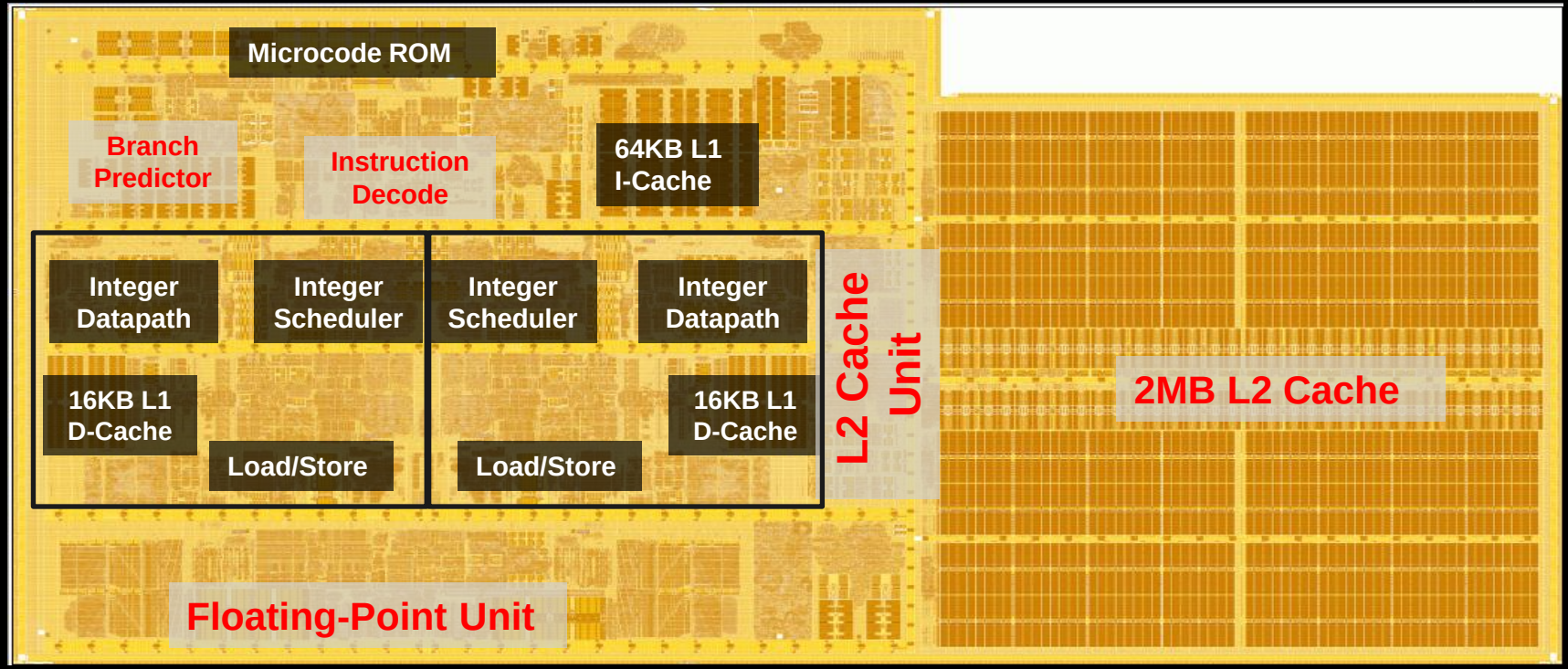
32nm "PILEDRIVER" COMPUTE MODULE

x86 CORE REDESIGN

- **Shared Fetcher / prediction pipeline - 64KB I-Cache**
- **Shared 4-way x86 decoder**
- **Shared Floating Point Unit - dual 128-bit FMA pipes**
- **Shared 16-way 2MB L2;**
- **Dedicated integer cores**
 - Register renaming based on physical register file
 - Unified scheduler per core
 - Way-predicted 16KB L1 D-cache
 - Out-of-order Load-Store Unit
- **ISA additions: FMA3, F16C**
- **Lightweight profiling support in HW**
- **"Piledriver" performance increase over "Stars"**
 - 14% improvement for desktop[5]
 - 25% improvement for notebook[2]
 - AMD Turbo Core 3.0



“PILEDRIVER” CORE FLOOR PLAN



"PILEDRIVER" IMPROVEMENTS & ENHANCEMENTS VS. "BULLDOZER"

"Bulldozer" Hybrid Predictor Augmented with 2nd level predictor

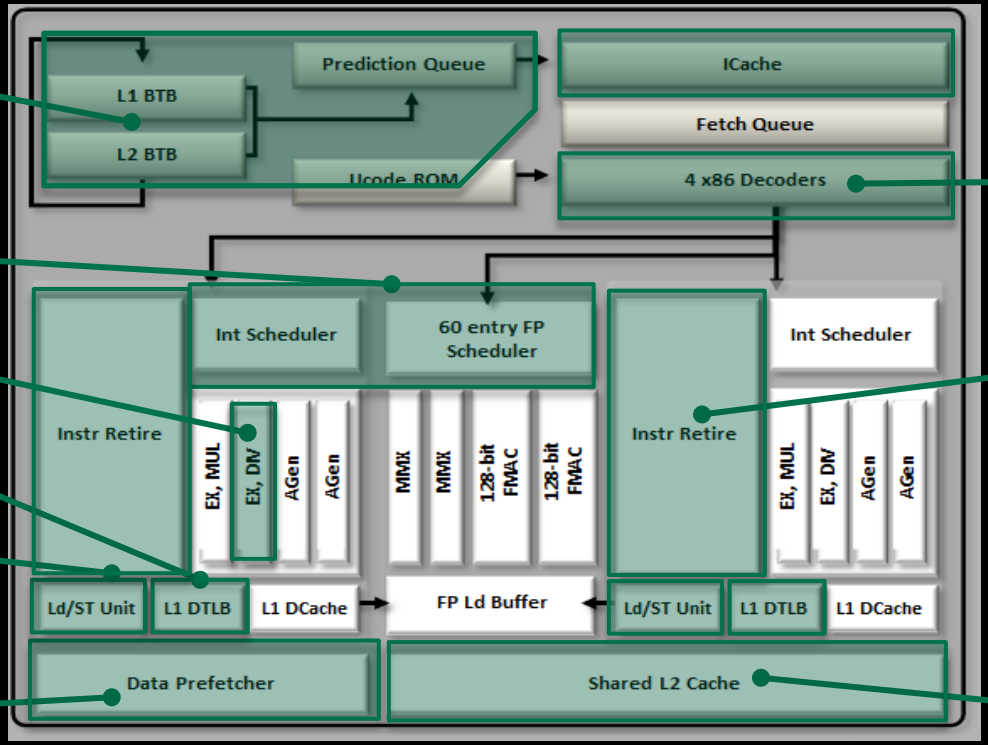
Improved scheduling FPU and INT

HW Divider

2x Larger L1 TLB

Improved Store-to-Load Forwarding

HW L1 Pre-fetcher improvements



ISA extensions FMA3, F16C

Faster Instruction exe SYSCALL & SYSRET

L2 efficiency and prefetching improvements

"PILEDRIVER" IMPROVEMENTS

30% higher CPU Freq [13]

- Design optimized for wide operational range (0.8V to 1.3V)
- 30% higher frequency at same voltage as "Stars" CPU Core in "Llano"

10% lower dynamic power vs. "Bulldozer"[13]

- Loop Predictor
- Way Predictor
- Dispatch gating based on group size
- Clock Gating
- Reduction in high power flops

Efficient operation

- 50% more base product frequency vs. "Llano" at same 35W SOC TDP

A10-4600M vs A8-3600M

Power management Latency reduction

- Intelligent L2 content tracking to speed up L2 flush
- State save/restore latency improvements to speed up power gating

MEDIA PROCESSING ACCELERATION



AMD'S UNIFIED VIDEO DECODER (UVD)



"TRINITY" ACCELERATED VIDEO CONVERTER ("AVC")

Core functionality

- Multi-stream hardware H.264 HD Encoder
- Power-efficient and faster than real-time[6] 1080p @60fps

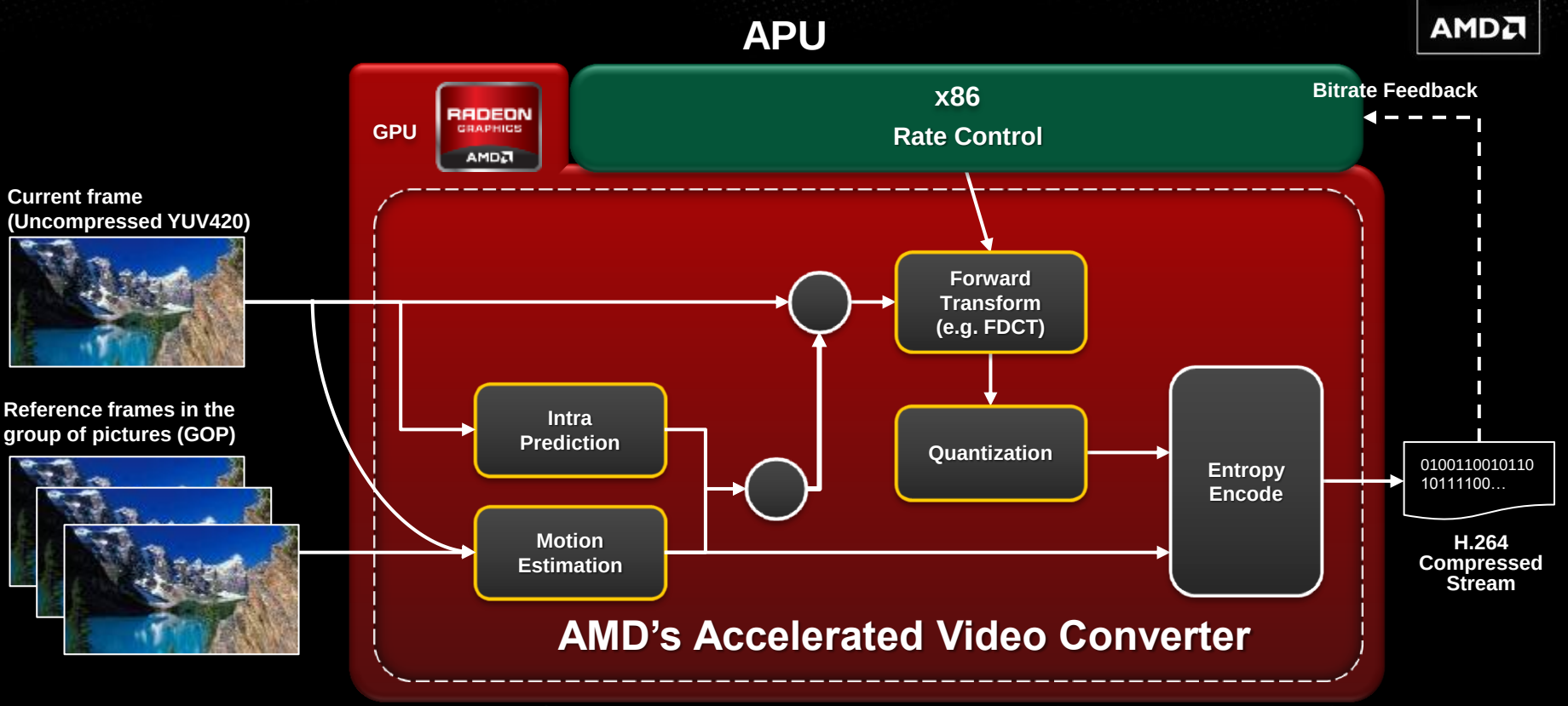
Quality features

- 4:2:0 color sampling video
- Optimizations for scene changes (games and video)
- Variable compression quality

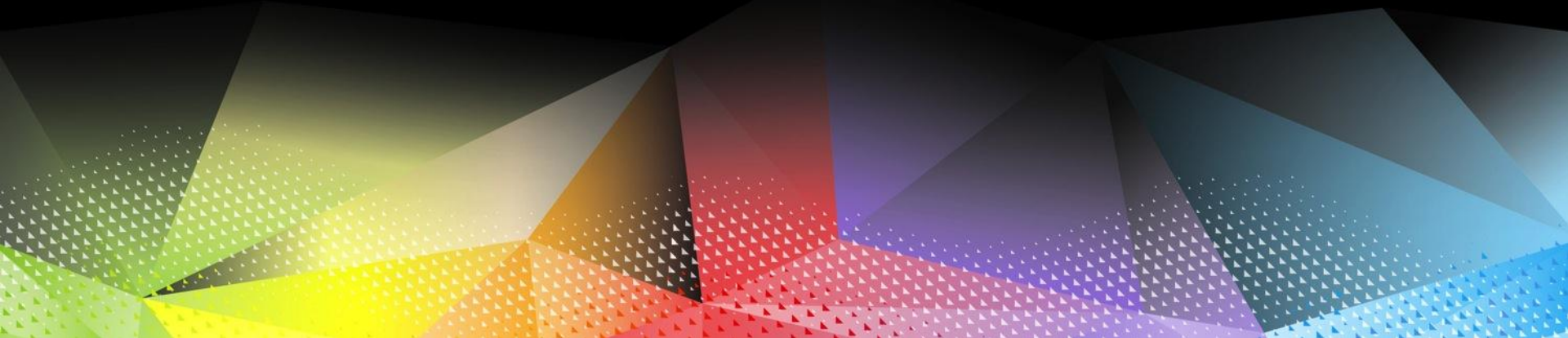
Interfacing features

- Audio / Video multiplexing
- Input from frame buffer for transcoding and video conferencing
- Input from GPU display engine for wireless display[7]

VIDEO ENCODING SYSTEM OPERATION

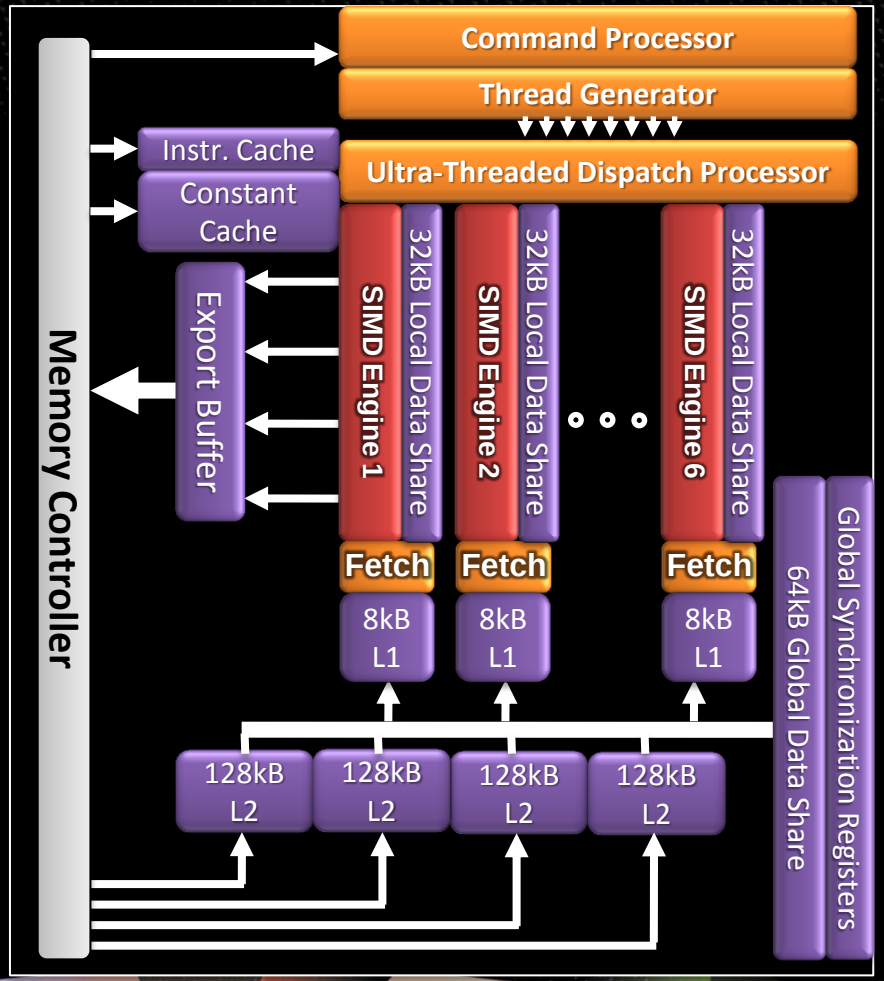


GPU DESIGN UPDATES FOR GAMING AND COMPUTE



3D ENGINE

- **DirectX® 11 – SM 5.0, OpenCL™ 1.1, DC 11**
- **GPU Core made of 384 Radeon™ Cores , each capable of 1 SP FMAC per cycle**
 - Organized as 96 stream processing units – each 4-way VLIW (vs. 5-way in Llano)
 - 6 SIMDs (each contains 16 processing units)
 - Each SIMD share 1 texture unit – achieving 4:1 ALU:Texture rate
- **32 depth / stencil per clock, 8 color per clock**
- **24x multi-sample and super sample, 16x anisotropic filtering**
- **Improved hardware tessellator vs. "Llano"**
- **Compute improvements**
 - Asynchronous dispatch: multiple compute kernels with independent address space simultaneously

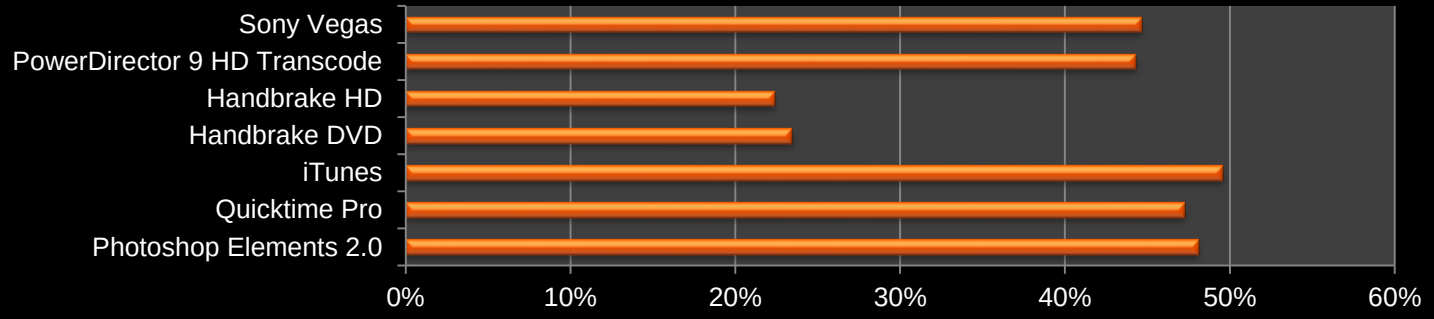


PERFORMANCE ACHIEVEMENTS

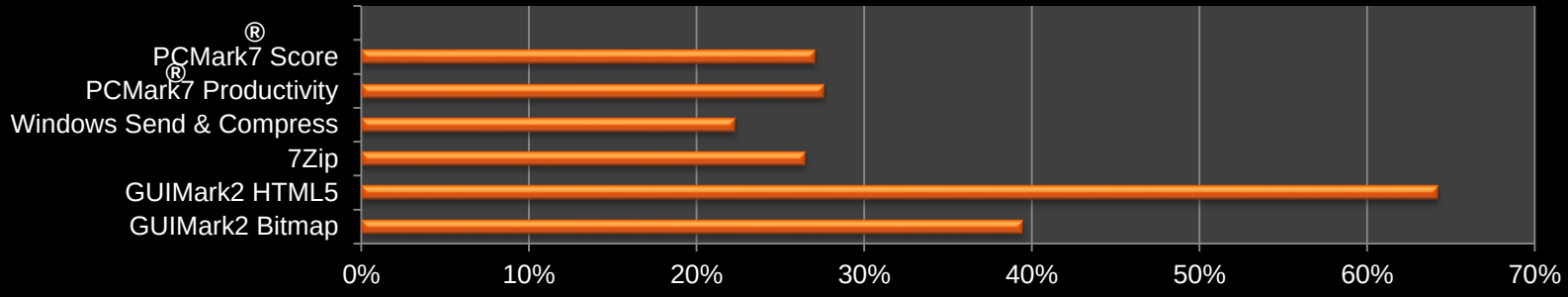
PERFORMANCE INCREASE ON CLIENT WORKLOADS (FOR 35W TDP)

"TRINITY" VS. "LLANO" CPU PERFORMANCE INCLUDING POWER MANAGEMENT, FREQ AND IPC GAINS

Digital Media

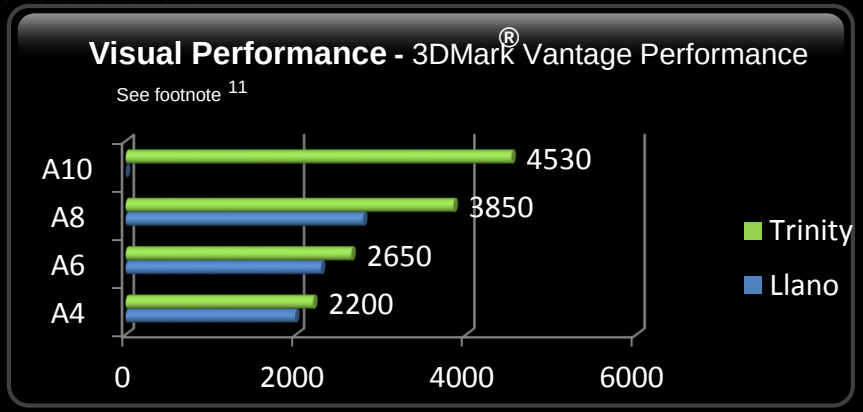


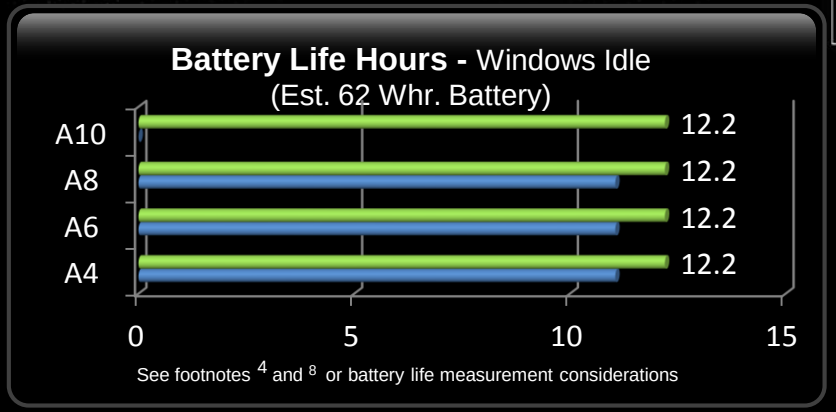
Web & Productivity: Compression & Cryptography

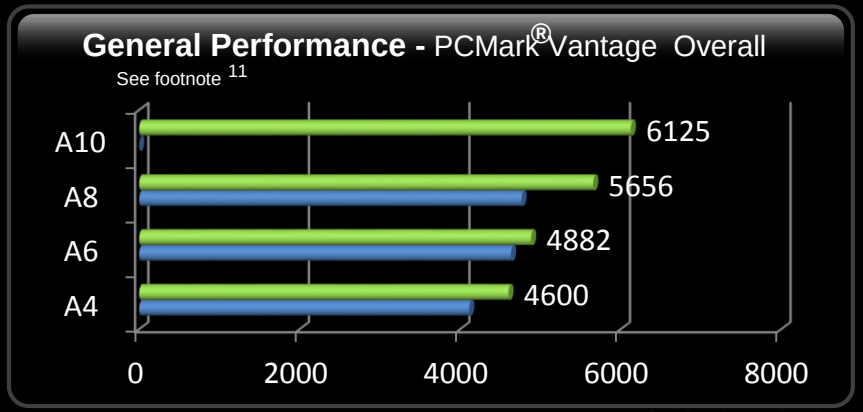


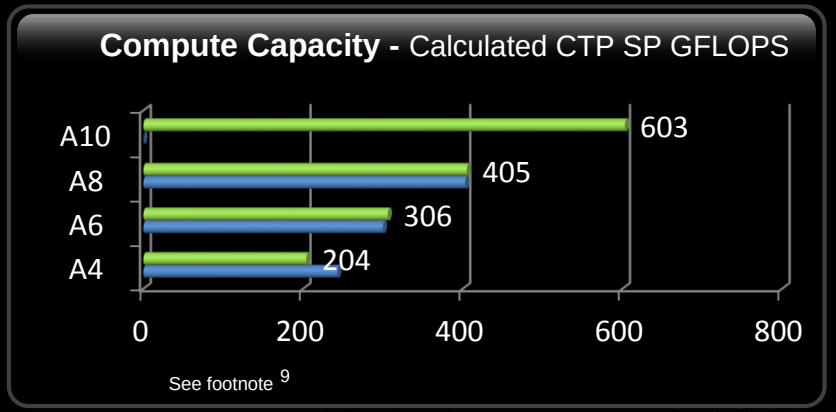
Experimental setup: see footnote 10

PERFORMANCE AND POWER COMPARISON VS PRIOR-GENERATION









Trinity performance based on estimates and/or preliminary benchmarks and are subject to change.

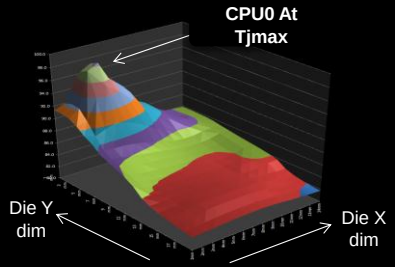
AMD TURBO CORE 3.0 TECHNOLOGY

AMD TURBO CORE 3.0 TECHNOLOGY : OVERVIEW

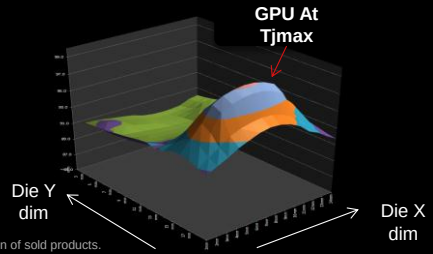
UTILIZE CALCULATED AVAILABLE DYNAMIC THERMAL HEADROOM TO IMPROVE PERFORMANCE

- **10-20 ºC variations across die during peak load**

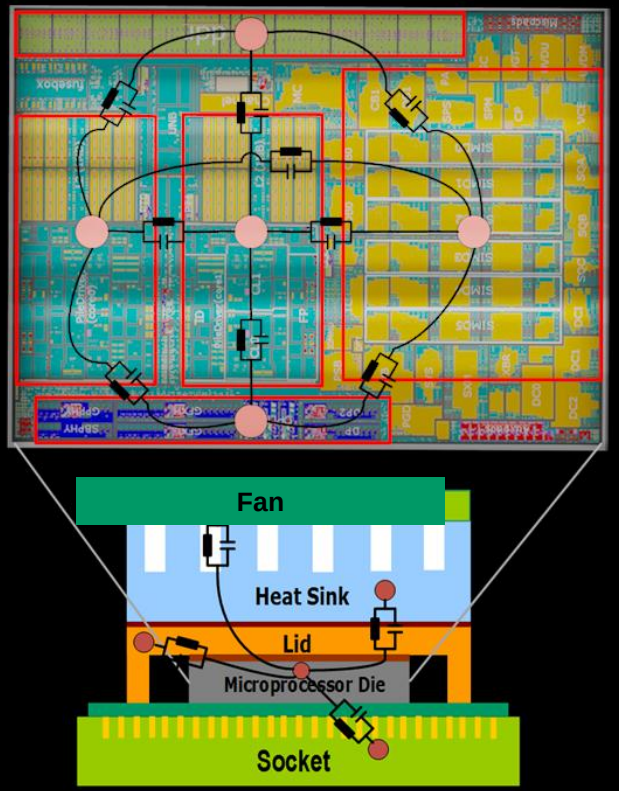
CPU-dominated workload (Livermore Loop 1Thread)
CPU0=17W, GPU=4.2W



GPU-dominated workload (3DMark®) with single thread application on CPU
CPU0=2.7W, GPU=23.9W

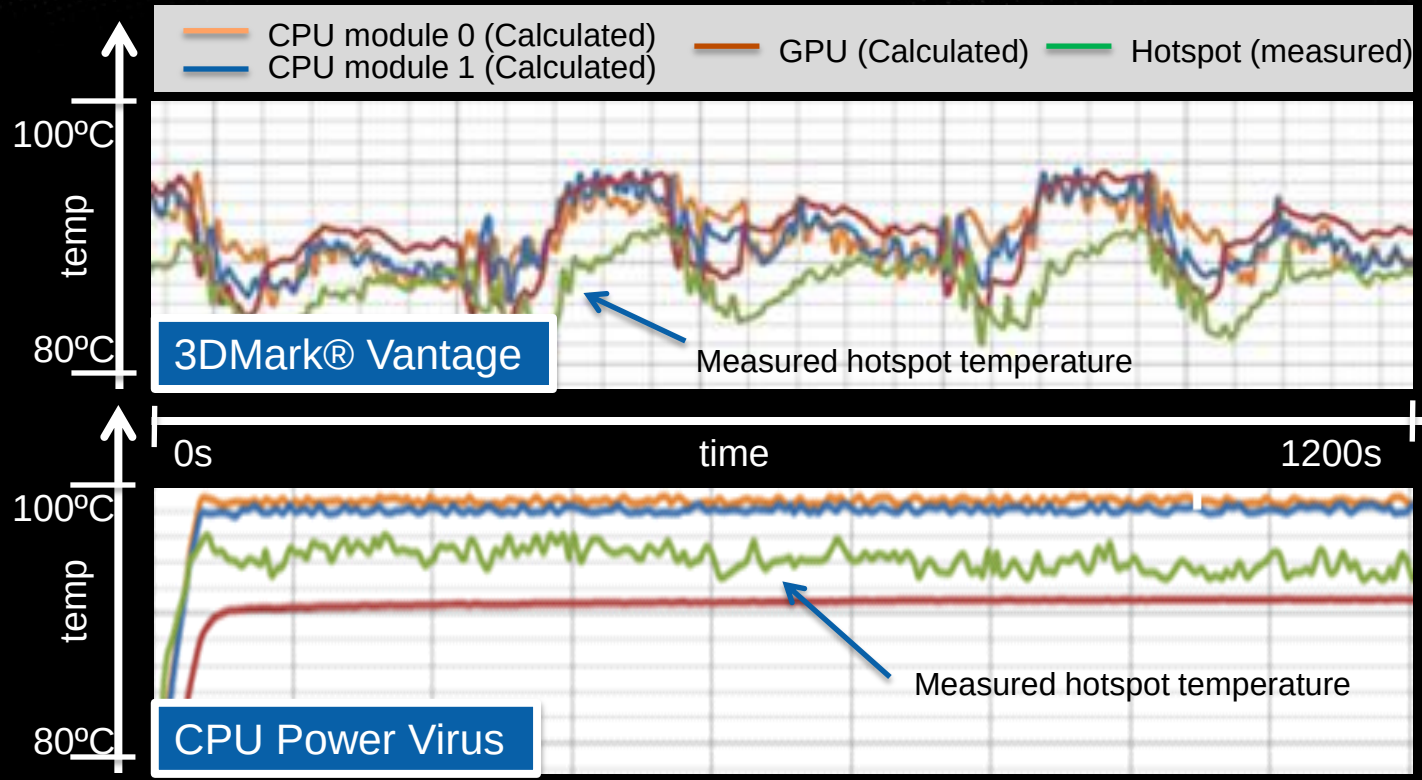


Simulation results for engineering discussion – no claims made to applicability to specific configuration of sold products.



- **Chip divided into "Thermal Entities" (TE)**
 - Thermal Entity calculate power and thermal density
- **Thermal RC network**
 - Transfer coefficients that describe thermal transfer between TEs, substrate and package are characterized
 - Numerical analysis firmware runs on the management processor which calculates per TE temperatures
 - TEs are throttled using voltage/frequency adjustments according to workload heuristics

AMD TURBO CORE 3.0 TECHNOLOGY: CALCULATED VS. MEASURED TEMPERATURE

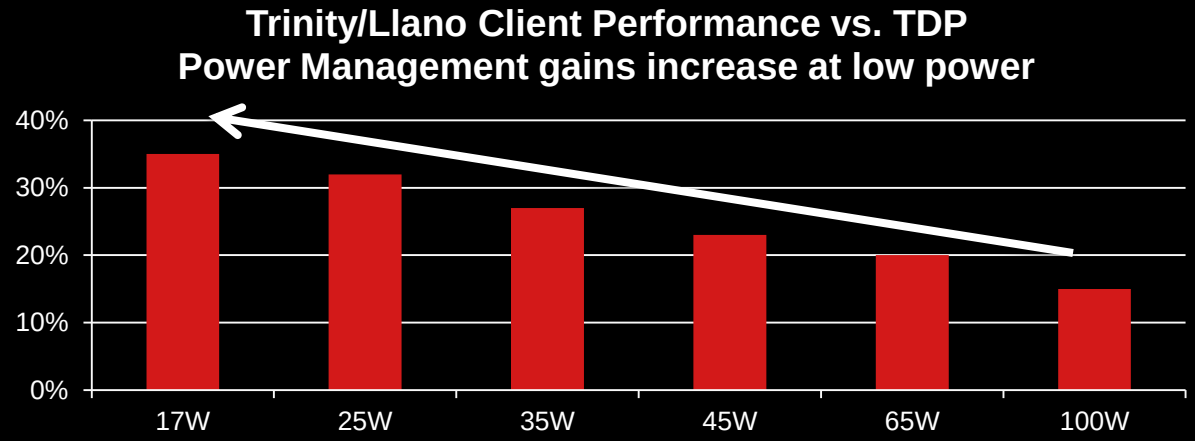


Estimated +/- 3-5C difference in calculated hotspot vs. measured hot spot temperature, at steady thermal state

Experimental results for engineering review, no observable product functional operational difference results from thermal differences. No claims made to accuracy

AMD TURBO CORE 3.0 TECHNOLOGY – PERFORMANCE

- **Workloads of moderate activity have high residency at maximum frequency**
 - Thermal headroom allows hotspot to remain below maximum control temperature
- **Higher activity workloads offer fewer opportunities to raise frequency and benefit from intelligent algorithms to bias power levels between CPU and GPU**
 - Collaborative or compute CPU/GPU applications
 - Multi-threaded workloads



Setup information: see footnote 12

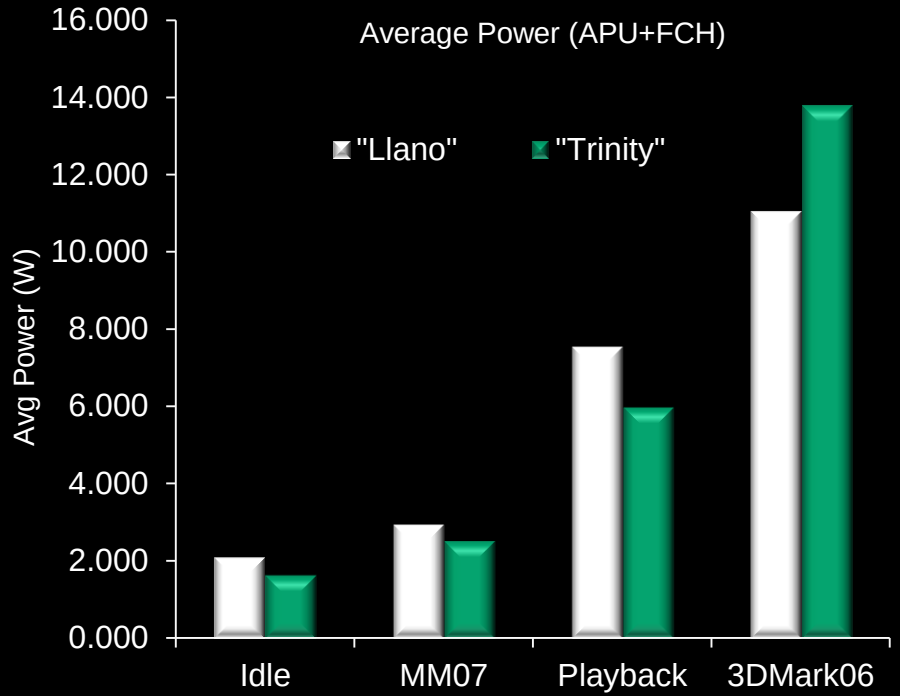
HARDWARE IMPROVEMENTS FOR LOW POWER



SYSTEM OPTIMIZATION POINTS

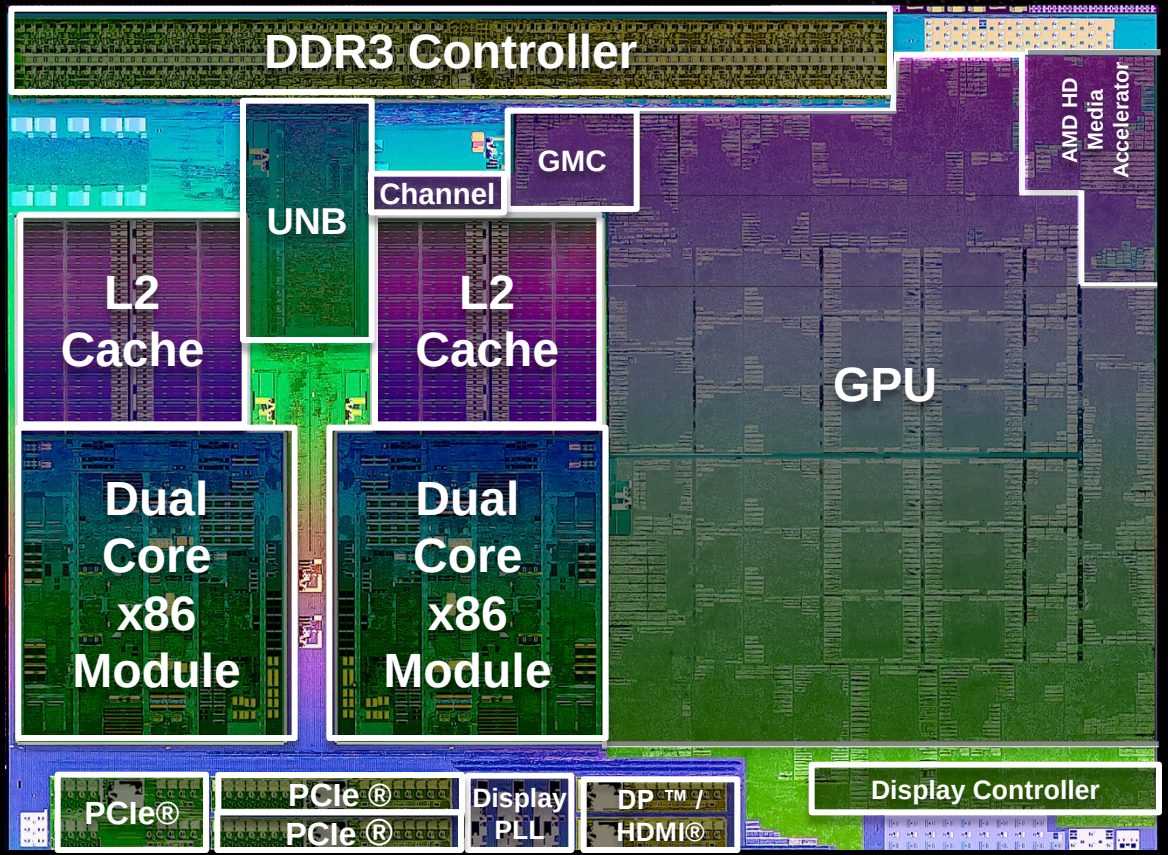
LEAP IN LOW POWER DESIGN

- **Idle — blank screen — system on**
- **MM07 — Mobile Mark 07**
- **Media playback — user experience**
- **Performance computing / gaming**
 - "Trinity" increases performance within fixed cooling solution
 - Trinity's significantly higher performance results in lower energy consumed for fixed amount of work or frames rendered, but higher power consumption during work

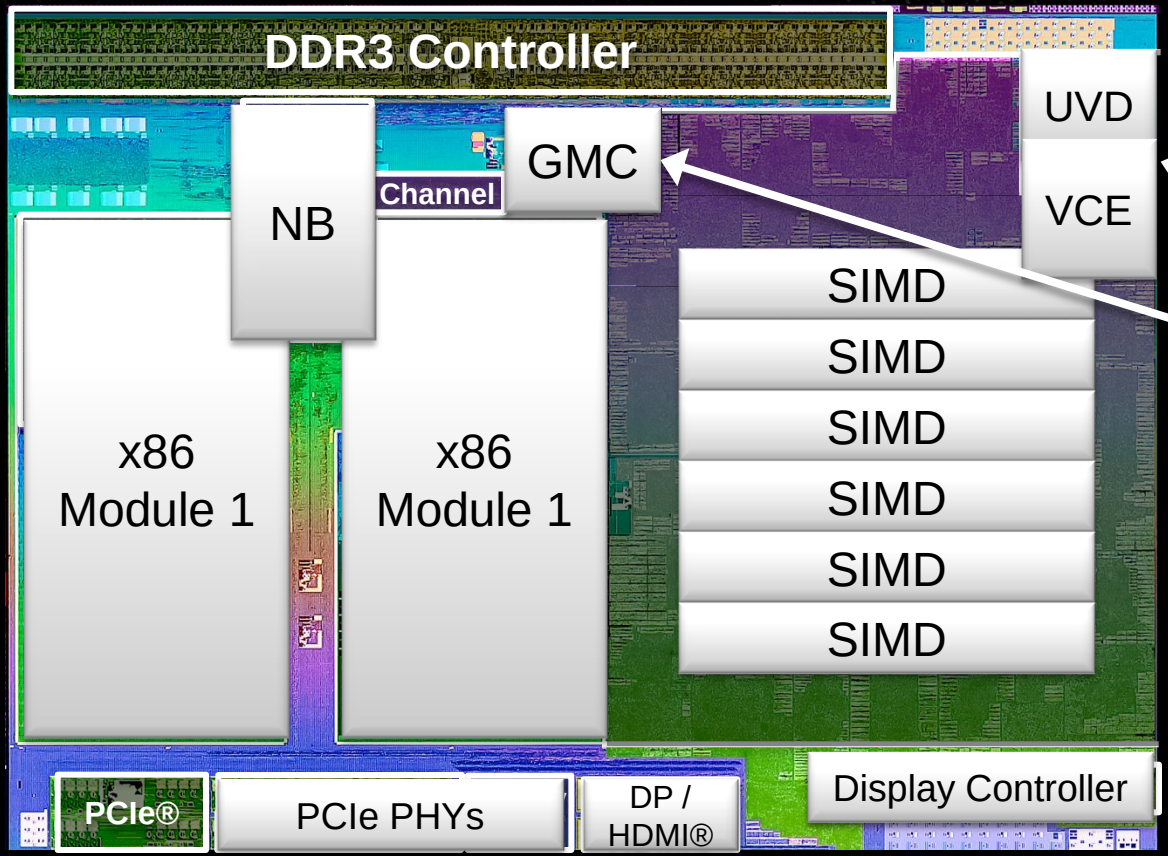


AMD A10-4600M APU on AMD "Pumori" reference board, 2x2GB DDR31600, SSD C300, Windows ® 7 64bit. Catalyst™ 8.941 vs A8-3600M,. 2x2GB DDR31600, SSD C300, Windows 7 64bit. Driver 8.941.Testing done at 1366x768. See footnotes [4] and [8] or battery life measurement considerations.

“TRINITY” APU FINE-GRAIN POWER GATING ISLANDS



"TRINITY" FINE-GRAIN POWER GATING ISLANDS (2)

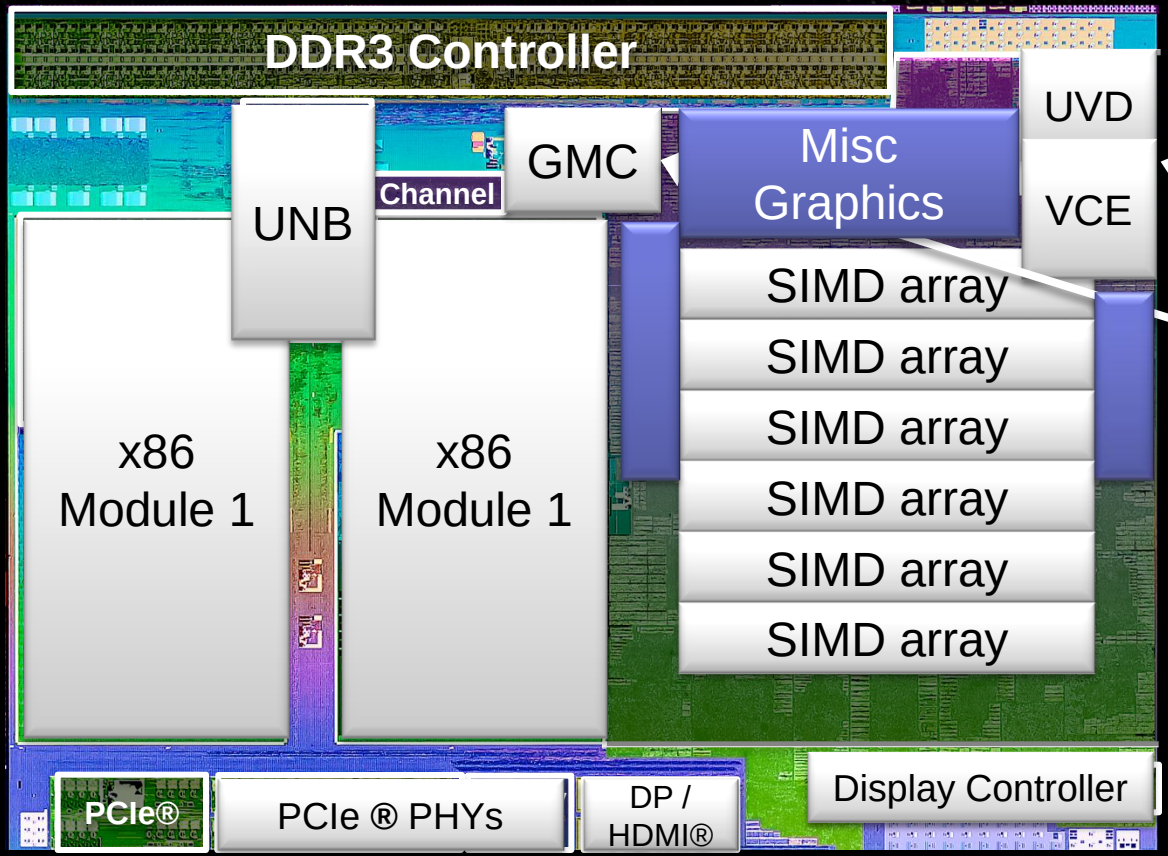


AMD HD Media Accelerator (UVD, AMD Accelerated Video Converter)

Graphics Memory Controller (Graphics optimized memory request scheduler)

Independent
on-die power gated islands

"TRINITY" FINE-GRAIN POWER GATING ISLANDS (3)



AMD HD Media Accelerator (UVD, AMD Accelerated Video Converter)

Graphics Memory Controller (Graphics optimized memory request scheduler)

Independent on-die power gated islands

Additional power-down region when all graphics functions are shut down

SMART NORTHBRIDGE OPERATIONAL STATES

- **Goals**
 - Intelligent selection of DRAM and Northbridge frequency to meet performance and power needs
 - Additional power savings from reduced DDR termination and drive strengths at low DDR speeds
- **Design supports low-latency transitions between several operational V/F points**
 - 4 Northbridge frequencies, 2 DRAM clock rates
- **Intelligent frequency selection based on performance needs from CPU / GPU and Multi-media**
 - Memory intensive workloads and certain multi-media content types trigger switch to higher DDR speed
 - CPU intensive workloads switches to higher Northbridge frequency to improve latency to memory
 - Multi-media buffers store real-time data during low-latency switching (less than 10 µs)
- **Frequency selection is further optimized by**
 - Static user policy selection of battery or performance optimization
 - OS power management hints
 - Heuristics to ensure higher voltage and frequency will not result in additional work throttling

THANK YOU !

AMD “TRINITY” APU

- **Core redesign for greater Performance**
- **Audio and Video enhancements for the best media experience**
- **Improved GPU performance with Radeon™ Cores 2.0**
- **Low Power Leadership**




AMD
THANK YOU

FOOTNOTES

1. Testing performed by AMD Performance Labs. The score for the 2012 AMD A10-4600M on the “Pumori” reference design for PC Mark ® Vantage Productivity benchmark shows an increase of up to 29% over the 2011 AMD A8-3500M on the “Torpedo” reference design. The AMD A10-4600M APU has a score of 6125 and the 2011 AMD A8-3500M APU scored 4764.

2. Projections and testing developed by AMD Performance Labs. Projected score for the 2012 AMD Mainstream Notebook Platform “Comal” the “Pumori” reference design for 3D Mark ® Vantage Performance benchmark is projected to increase by up to 50% over actual scores from the 2011 AMD Mainstream Notebook Platform “Sabine.” Projections were based on AMD A8/A6/A4 35w APUs for both platforms.

3. AMD Eyefinity technology works with games that support non-standard aspect ratios, which is required for spanning across multiple displays. To enable more than two displays, additional panels with native DisplayPort™ connectors, and/or DisplayPort™ compliant active adapters to convert your monitor’s native input to your cards DisplayPort™ or Mini-DisplayPort™ connector(s), are required. AMD Eyefinity technology can support up to 6 displays using a single enabled AMD Radeon™ GPU with Windows Vista® or Windows® 7 operating systems .

4. Testing and calculations by AMD Performance Labs. Battery life calculations based on average power on multiple benchmarks and usage scenarios. These include Active metric using FutureMark® 3DMark ’06 (172 min./2:54 hours), streaming YouTube video (271 min./4:30 hours), playback of a Microsoft sample clip from local HDD (303 min./5:03 hours), PowerMark ® Productivity benchmark/radio off (483 min./8:03 hours), web browsing test was average of 40 minutes via 802.11n WLAN, 2 minutes per page using the web test tool developed by AMD (570 min./9:30 hours) and Windows ® Idle (725 min./12:05 hours) as a resting metric. All battery life calculations are based on using a 6 cell Li-Ion 62.16Whr battery pack at 98% utilization for Windows ® Idle, PowerMark ® and 96% utilization for 3DMark ® 06 workload, video playback and YouTube video streaming; and 92% utilization for Blu-ray playback.

5. Projections and testing developed by AMD Performance Labs. The AMD A-10 5800K APU with AMD Radeon™ HD 7660D graphics, versus an AMD A8-3850 APU with 14% uplift on x86 performance in measure in PCMark7 ® Productivity, and 30% planned uplift on graphics performance using 3DMark ® 11 (P). All systems using “Trinity” 100W APU, 8GB DDR3-16000 memory, Windows ®7 64 bit.

FOOTNOTES (2)

6. Based on AMD internal testing of video encoding speed of VCE of 1080p H.264 video at 47 seconds, which is faster than the 65 second size of the 480p-kid.mov video file. System configuration: OS: Windows ® 7 64-bit, CPU: AMD A10-5800K with AMD Radeon™ HD 7660D graphics, Annapurna reference board, 8GB DDR3-1600, Windows ® 7 64bit.

7. AMD Wireless Display technology provides the ability to wirelessly display local screen content onto a remote screen in real time. Compliant receiver equipment required.

8. Testing and projections conducted by AMD performance labs. Testing on the 2011 AMD Mainstream Notebook Platform show 663 minutes (11.05 hours) of Windows ® Idle as "resting" battery life. Projections for the 2012 AMD Mainstream Platform "Comal" show 748 minutes (12.47 hours) of Windows ® Idle as "resting" battery life.

9. GFLOPs calculations developed by AMD performance labs measuring compute capacity for the 2012 VISION A10-based notebook which scored 603 GFLOPS. AMD GFLOPs calculated using GFLOPs = CPU GFLOPs + GPU GFLOPs = CPU Core Freq. X Core Count X 8 FLOPS + GPU Core Freq. X DirectX® 11 capable Shader Count X 2 FLOPS.

10. Experimental results on A10-4600M with Radeon HD7660G Graphics ("Trinity") vs. A8-3500M 4GB DDR3-1600 with Radeon HD6620G Graphics ("Llano") 4GB DDR3-1333 – running under Windows ® 7 Ultimate, with Hitachi HDD 5400RPM

11. Projections and testing developed by AMD Performance Labs. Projected scores for the 2012 AMD Mainstream Notebook Platform "Comal" the "Pumori" reference design for 3D Mark ® Vantage Performance, PCMark ® Vantage over actual scores from the 2011 AMD Mainstream Notebook Platform "Sabine". Projections were based on AMD A10/A8/A6/A4 35w APUs.

12. AMD A10-4600M APU with Radeon(tm) HD Graphics, 4GB DDR3-1600, on Pumori Reference Board with Hitachi 5400 RPM HDD.

13. Power measured by AMD Perf Labs on "Trinity" A0 silicon running SpecInt ® 2006 on Pumori Reference board, and on Orochi B0 (which contains "Bulldozer" Core) at same voltage and frequency. 20% dynamic power improvement was offsetted for caching structures differences and leads to an estimate of more than 10% dynamic power reduction directly attributable to the Core ® on SpecInt 2006. Frequency improvement vs. "Stars" Core measured by AMD PEO for nominal process targeting on "Llano" Rev. B0 and "Trinity" Rev. A1

DISCLAIMER & ATTRIBUTION



The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.

The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. There is no obligation to update or otherwise correct or revise this information. However, we reserve the right to revise this information and to make changes from time to time to the content hereof without obligation to notify any person of such revisions or changes.

NO REPRESENTATIONS OR WARRANTIES ARE MADE WITH RESPECT TO THE CONTENTS HEREOF AND NO RESPONSIBILITY IS ASSUMED FOR ANY INACCURACIES, ERRORS OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION.

ALL IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE ARE EXPRESSLY DISCLAIMED. IN NO EVENT WILL ANY LIABILITY TO ANY PERSON BE INCURRED FOR ANY DIRECT, INDIRECT, SPECIAL OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

AMD, the AMD arrow logo and combinations thereof are trademarks of Advanced Micro Devices, Inc. All other names used in this presentation are for informational purposes only and may be trademarks of their respective owners.

AMD, the AMD Arrow logo and combinations thereof are trademarks of Advanced Micro Devices, Inc. in the United States and/or other jurisdictions. Other names used in this presentation are for identification purposes only and may be trademarks of their respective owners.

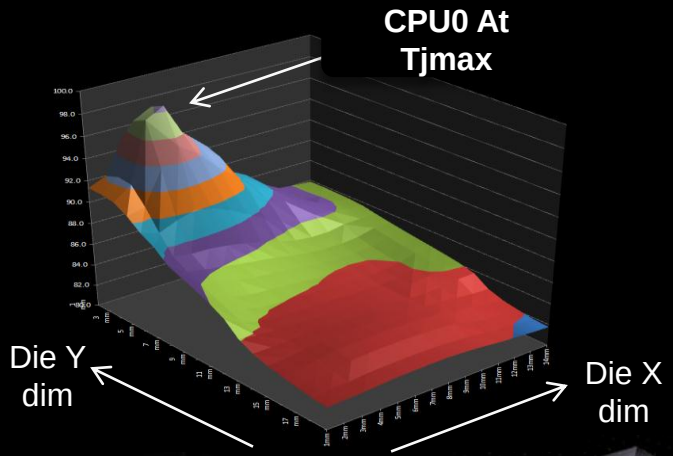
© 2012 Advanced Micro Devices, Inc. All rights reserved.

APPENDIX

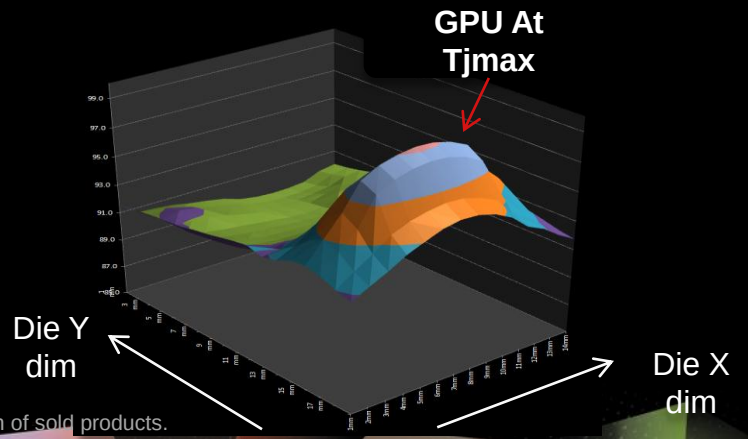
AMD TURBO CORE 3.0 TECHNOLOGY – DIE THERMAL SIMULATIONS

- **Thermal simulations for a 35W product**
 - 10-20ºC variations across the die depending on the workload, during peak activity
 - Hotspot needs to be controlled to maximum junction temperature
 - Hotspot thermal simulations are now critical part of the performance optimization flow

CPU-dominated workload (Livermore Loop 1Thread)
CPU0=17W, GPU=4.2W



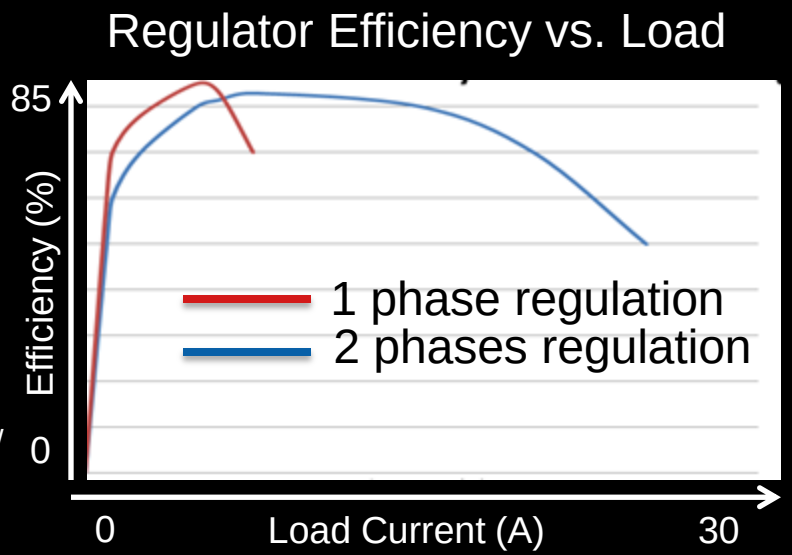
GPU-dominated workload (3DMark®) with single thread application on CPU
CPU0=2.7W, GPU=23.9W



Simulation results for engineering discussion – no claims made to applicability to specific configuration of sold products.

PLATFORM POWER SAVINGS : AMD SERIAL VOLTAGE INTERFACE 2.0

- **SVI is the interface which allows the processor to communicate information to and from voltage regulator**
- **SVI2 enables quicker power state transitions**
 - Faster data transmission rates (33Mhz)
 - New regulator response when transition is complete
 - 80+% improvement in 500mV set point change latency
- **Power efficiency features**
 - Multiple Power State Indicators sent to regulator
 - PSI0 – Current low enough that regulator can shed phases
 - PSI1 – Current low enough that regulator can use pulse skipping / diode emulation
 - Load Line trim, offset
 - Ability to adjust DC offset and load line slope based on APU state



"TRINITY" DESIGN FOR LOW POWER – NEW FEATURES

Display Power Optimizations

- Static-screen display refresh from single DRAM channel
- On-die cursor caching
- Increased on-chip buffering of display memory

Power Tuning

- Voltage and frequency are automatically selected using indication from
 - GPU Power state, PCIe ® speed, Multimedia workload
- Dynamic DRAM speed – reduced power when bandwidth requirements are low
- SVI-2 Voltage Regulator interface –selection of optimal regulator power state depending on load

Power Gating, low voltage I/O

- UNB Power Gated when idle
- GPU Core support per-SIMD power gating
- PCI-Express® and Display PHY Power Gating
- Accelerated Video Converter Power down
- Support for 1.25V DDR3 Memory

Graphics and Multimedia

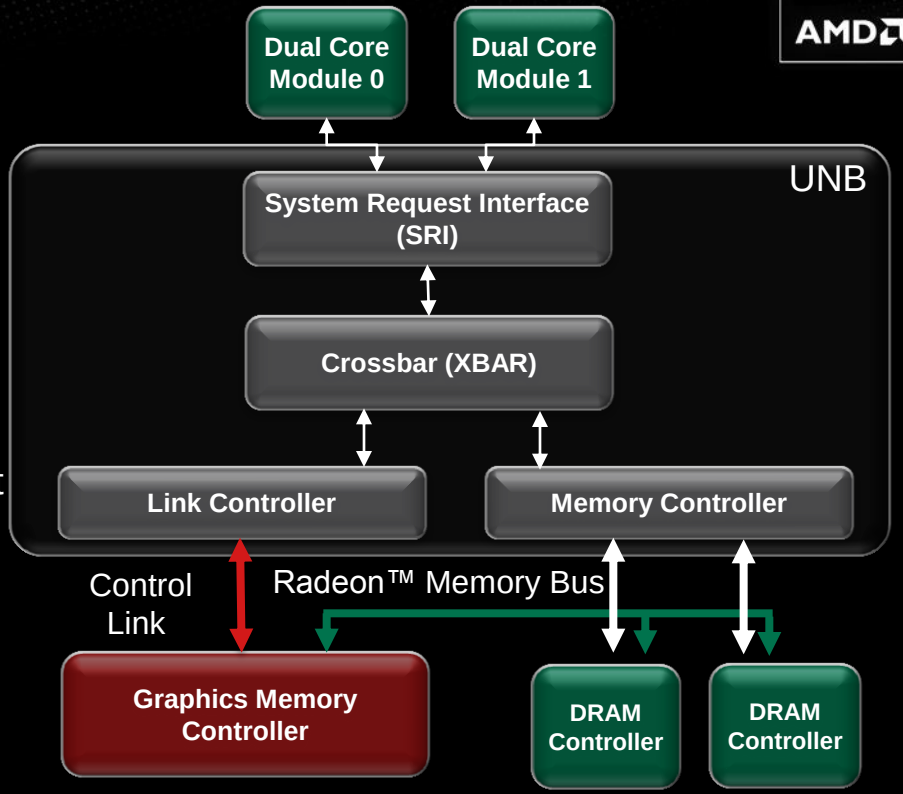
- Video Compression Engine – offload engine to save encoding power

AMD TURBO CORE 3.0 TECHNOLOGY : TEMPERATURE CALCULATION

- **CPU/GPU Temperature**
 - Firmware regularly calculates instantaneous temperature for each TE new power estimate and prior temperature
 - Uses a 5 stage thermal RC ladder
- **Other silicon contributors**
 - High Speed IO interfaces, Northbridge are modeled as power and/or temperature offsets to simplify calculations
 - This has limited impact on accuracy
- **Measured error of +/-5 ºC on 3DMark® analysis**
- **Algorithm provides deterministic operation and reproducibility of results**

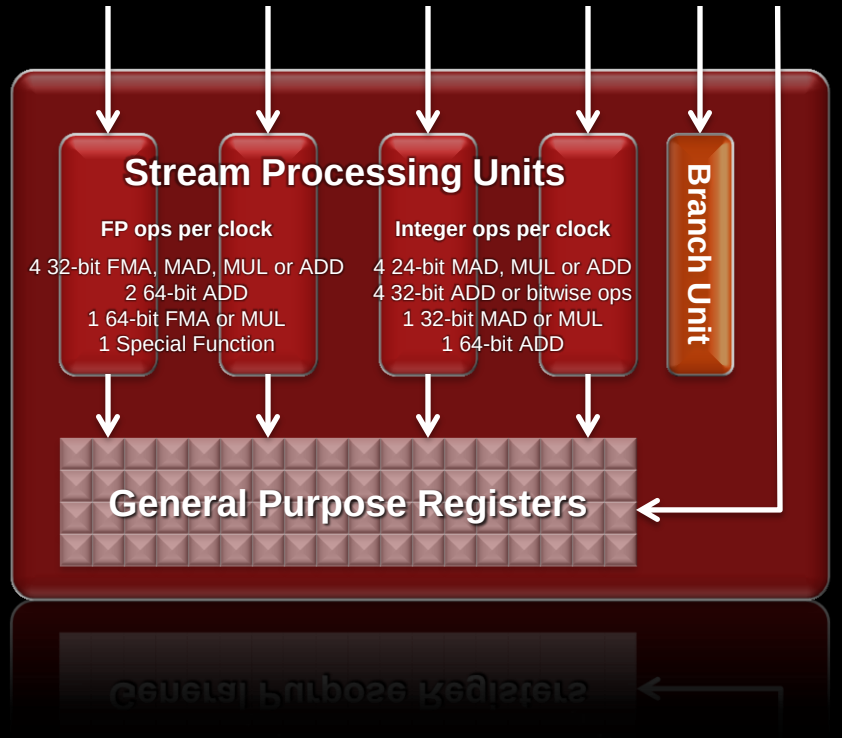
UNIFIED NORTHBRIDGE AND MEMORY

- **Unified Northbridge**
 - First UNB for APUs featured on “Trinity”
 - Supports:
 - Interface to a Graphics Memory Controller
 - Two DDR2/3 interfaces, shared with the Graphics Northbridge via Radeon Memory Bus
 - APU Power Management
- **Memory Support**
 - 128-bit interface arranged as two un-ganged 64-bit channels
 - Supports Memory P-states — with memory speed changes on the fly
 - Supports 1.25V DIMMS
 - Up to 29.8 GB/s with DDR3-1866



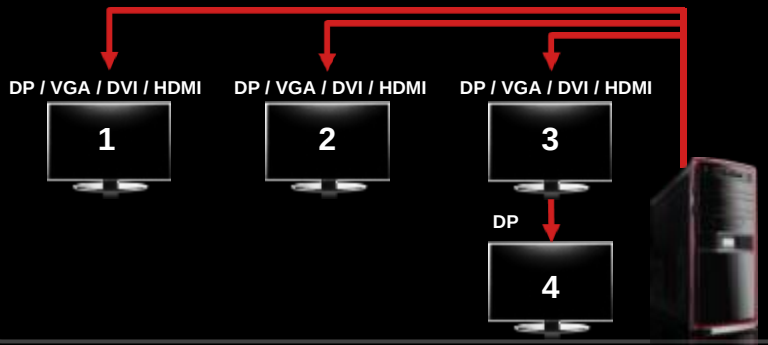
RADEON CORE S 2.0 DESIGN

- **VLIW4 thread processors**
 - 4-way co-issue
 - All stream processing units now have equal capabilities
 - Special functions (transcendentals) occupy 3 of 4 issue slots
- **Allow better utilization than previous VLIW5 design**
 - Improved performance/mm²
 - Simplified scheduling and register management
 - Extensive logic reuse

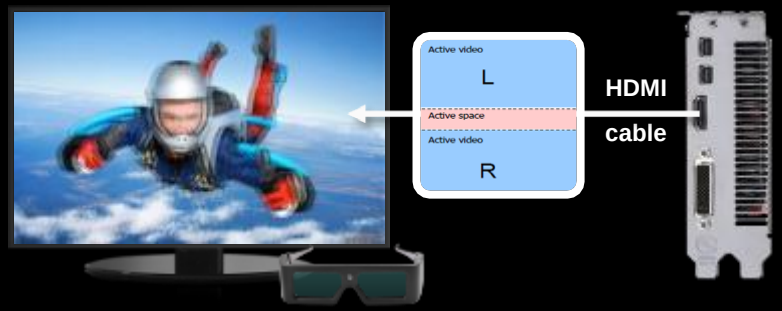


DISPLAY TECHNOLOGY LEADERSHIP

DP1.2



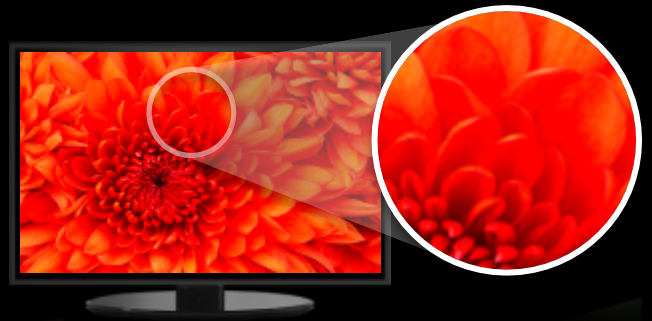
HDMI



Stereo 3D



Color Correction





Intel® Xeon Phi™ coprocessor (codename Knights Corner)

George Chrysos
Senior Principal Engineer
Hot Chips, August 28, 2012

Legal Disclaimers

Copyright © 2012 Intel Corporation. All rights reserved.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

A "Mission Critical Application" is any application in which failure of the Intel Product could result, directly or indirectly, in personal injury or death. SHOULD YOU PURCHASE OR USE INTEL'S PRODUCTS FOR ANY SUCH MISSION CRITICAL APPLICATION, YOU SHALL INDEMNIFY AND HOLD INTEL AND ITS SUBSIDIARIES, SUBCONTRACTORS AND AFFILIATES, AND THE DIRECTORS, OFFICERS, AND EMPLOYEES OF EACH, HARMLESS AGAINST ALL CLAIMS COSTS, DAMAGES, AND EXPENSES AND REASONABLE ATTORNEYS' FEES ARISING OUT OF, DIRECTLY OR INDIRECTLY, ANY CLAIM OF PRODUCT LIABILITY, PERSONAL INJURY, OR DEATH ARISING IN ANY WAY OUT OF SUCH MISSION CRITICAL APPLICATION, WHETHER OR NOT INTEL OR ITS SUBCONTRACTOR WAS NEGLIGENT IN THE DESIGN, MANUFACTURE, OR WARNING OF THE INTEL PRODUCT OR ANY OF ITS PARTS.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.
Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm%20

Intel, the Intel logo, Xeon, Intel Core and Intel Xeon Phi are trademarks of Intel Corporation in the U.S. and/or other countries Other names and brands may be claimed as the property of others.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

For more complete information about performance and benchmark results, visit Performance Test Disclosure

This document contains information on products in the design phase of development.

All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel.

Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Notice revision #20110804

WARNING: Altering clock frequency and/or voltage may: (i) reduce system stability and useful life of the system and processor; (ii) cause the processor and other system components to fail; (iii) cause reductions in system performance; (iv) cause additional heat or other damage; and (v) affect system data integrity. Intel has not tested, and does not warranty, the operation of the processor beyond its specifications. Intel assumes no responsibility that the processor, including if used with altered clock frequencies and/or voltages, will be fit for any particular purpose. For more information, visit Overclocking Intel Processors

Warning: Altering PC memory frequency and/or voltage may (i) reduce system stability and use life of the system, memory and processor; (ii) cause the processor and other system components to fail; (iii) cause reductions in system performance; (iv) cause additional heat or other damage; and (v) affect system data integrity. Intel assumes no responsibility that the memory, included if used with altered clock frequencies and/or voltages, will be fit for any particular purpose. Check with memory manufacturer for warranty and additional details

Available on select Intel® Core™ Intel® Xeon® and Intel® Xeon Phi™ processors. Requires an Intel® HT Technology-enabled system. Consult your PC manufacturer. Performance will vary depending on the specific hardware and software used. For more information including details on which processors support HT Technology, visit http://www.intel.com/info/hyperthreading.

Requires a system with a 64-bit enabled processor, chipset, BIOS and software. Performance will vary depending on the specific hardware and software you use. Consult your PC manufacturer for more information. For more information, visit http://www.intel.com/info/em64t

Requires a system with Intel® Turbo Boost Technology. Intel Turbo Boost Technology and Intel Turbo Boost Technology 2.0 are only available on select Intel® processors. Consult your PC manufacturer. Performance varies depending on hardware, software, and system configuration. For more information, visit http://www.intel.com/go/turbo

ENERGY STAR is a system-level energy specification, defined by the Environmental Protection Agency, that relies on all system components, such as processor, chipset, power supply, etc.) For more information, visit http://www.intel.com/technology/epa/index.html



Intel® Many Integrated Core (Intel MIC) Architecture

Targeted at highly parallel HPC workloads

- Physics, Chemistry, Biology, Financial Services

Power efficient cores, support for parallelism

- Cores: less speculation, threads, wider SIMD
- Scalability: high BW on die interconnect and memory

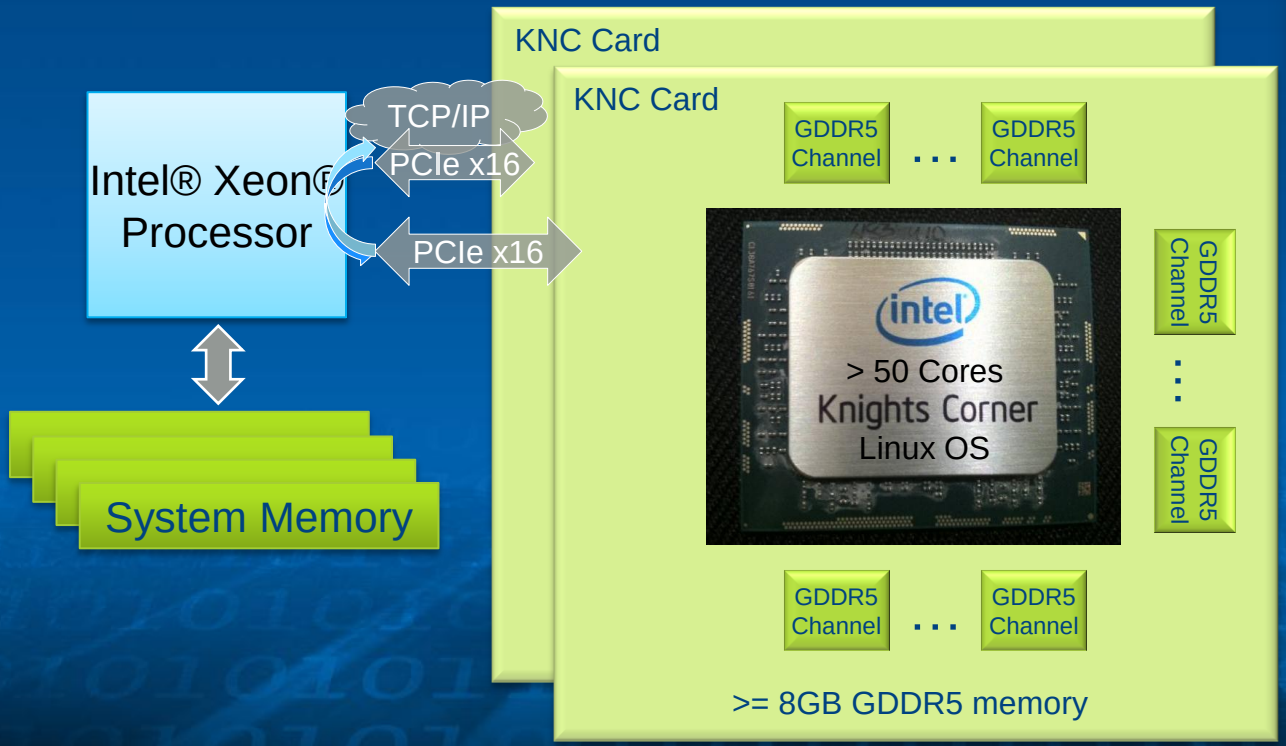
General Purpose Programming Environment

- Runs Linux (full service, open source OS)
- Runs applications written in Fortran, C, C++, …
- Supports X86 memory model, IEEE 754
- x86 collateral (libraries, compilers, Intel® VTune™ debuggers, etc)

Copyright © 2012 Intel Corporation. All rights reserved.

Knights Corner Coprocessor

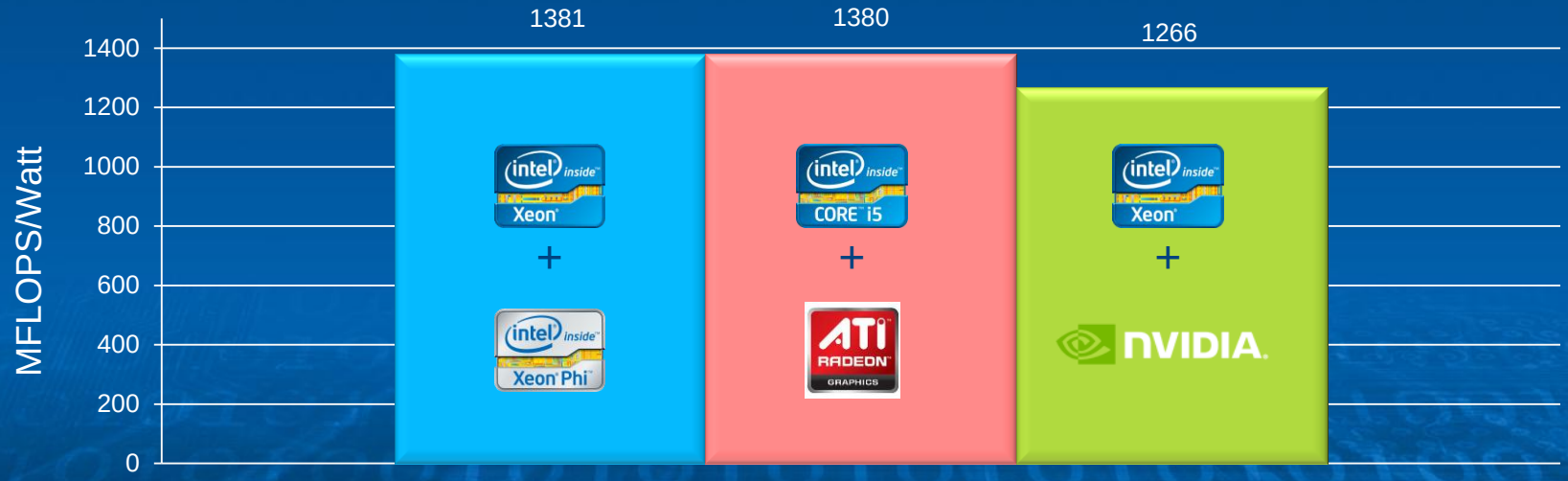




Copyright © 2012 Intel Corporation. All rights reserved.

Knights Corner – Power Efficient

Performance per Watt of a prototype Knights Corner Cluster compared to the 2 Top Graphics Accelerated Clusters

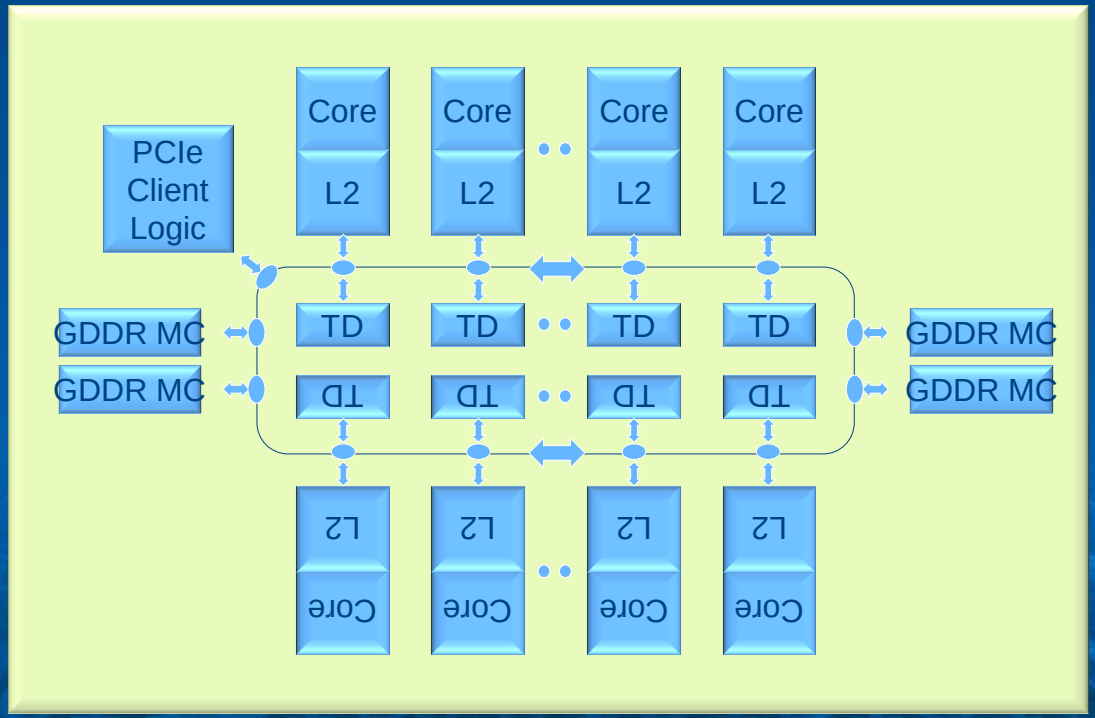


Higher is Better Source: www.green500.org

Intel Corp	Nagasaki Univ.	Barcelona Supercomputing Center
Knights Corner	ATI Radeon	Nvidia Tesla 2090
Top500 #150	Top500 #456	Top500 #177
72.9 kW	47 kW	81.5 kW

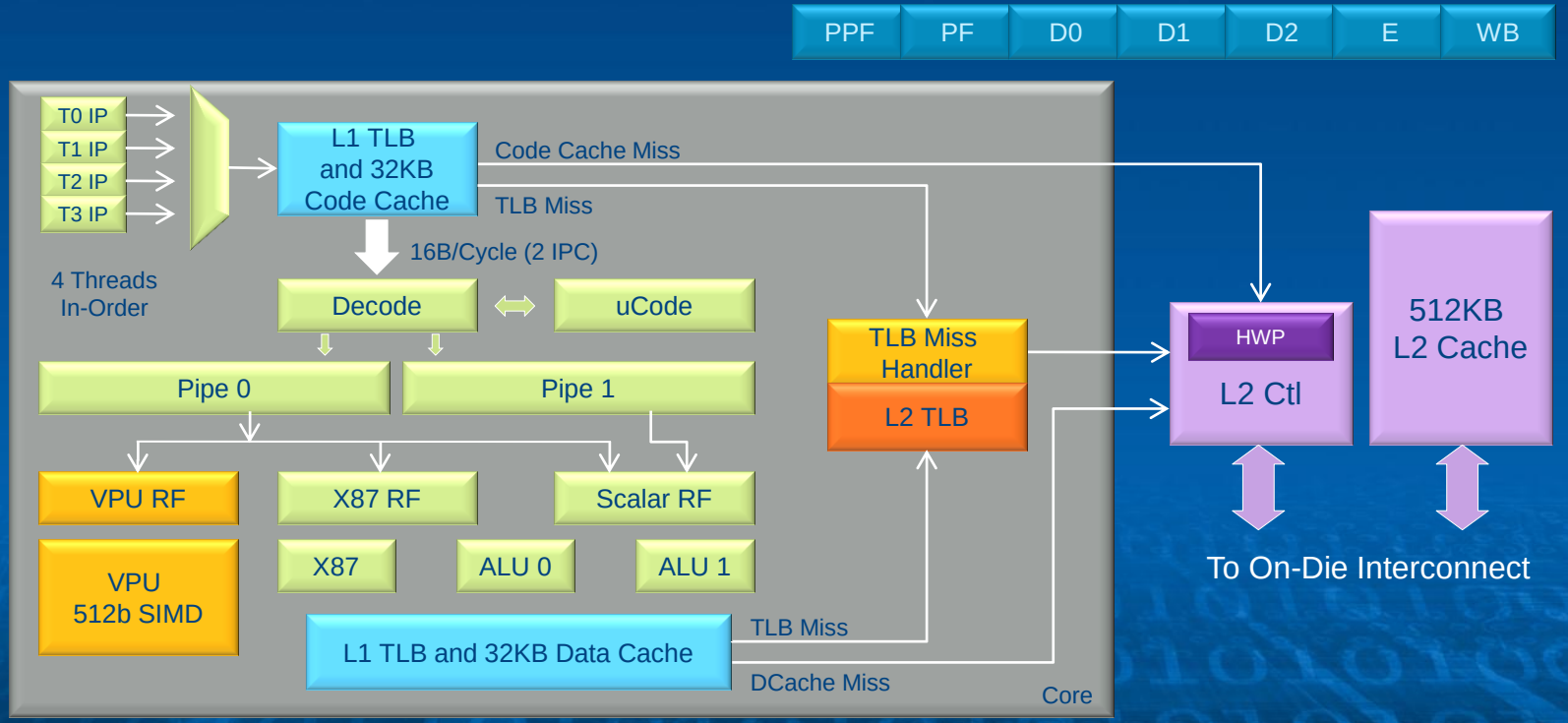
Copyright © 2012 Intel Corporation. All rights reserved.

Knights Corner Micro-architecture



Copyright © 2012 Intel Corporation. All rights reserved.

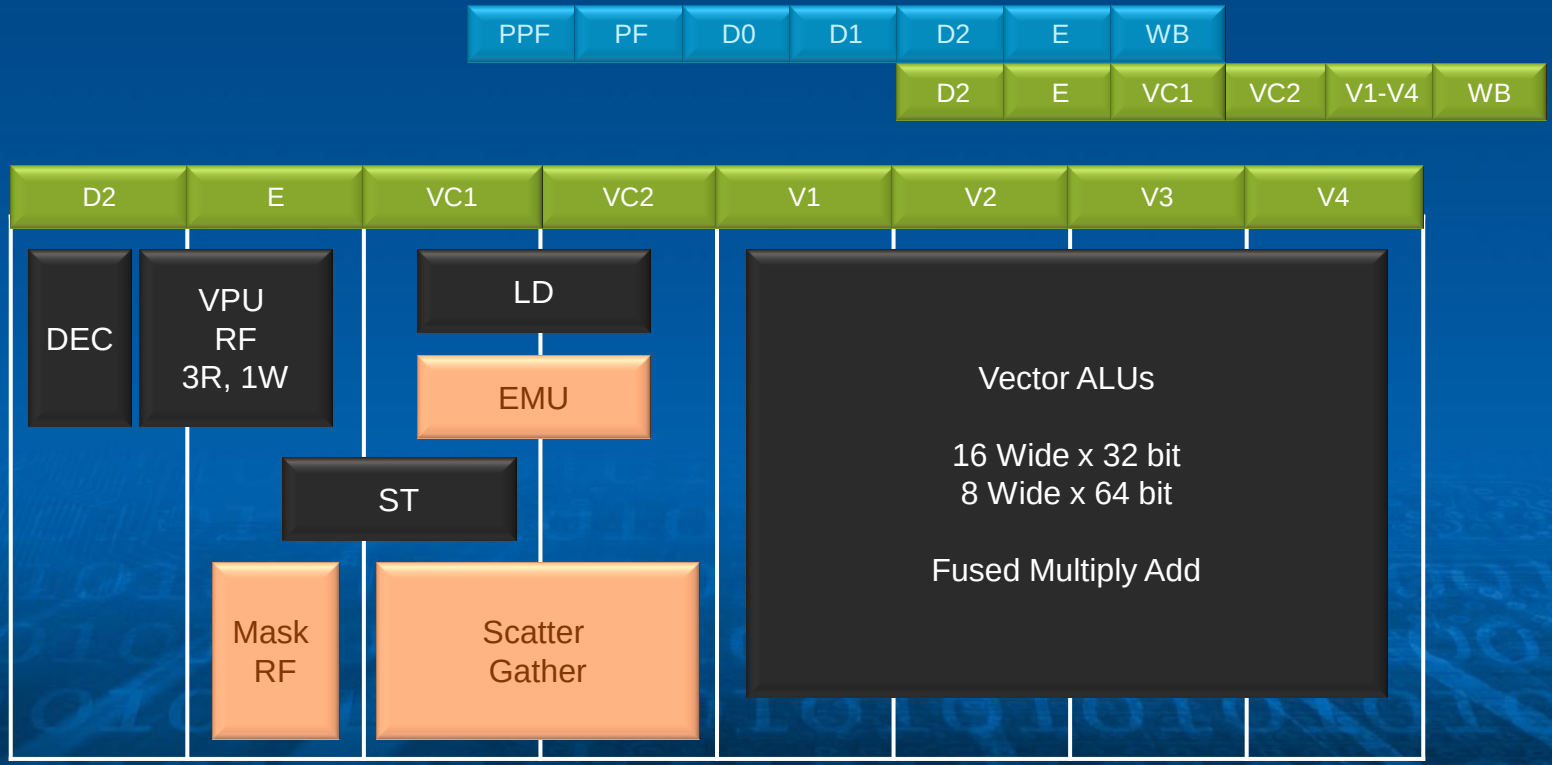
Knights Corner Core



X86 specific logic < 2% of core + L2 area

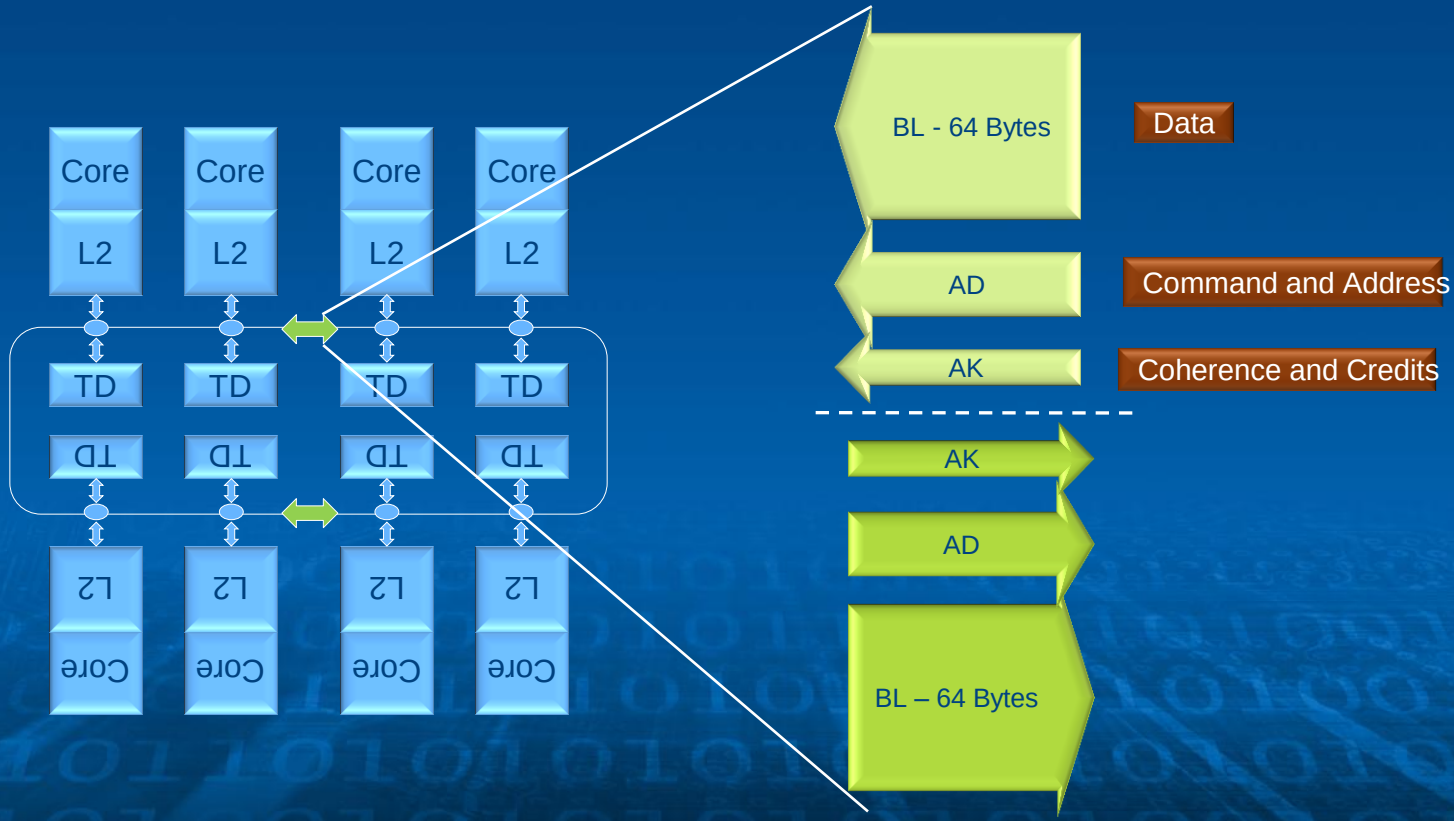
Copyright © 2012 Intel Corporation. All rights reserved.

Vector Processing Unit



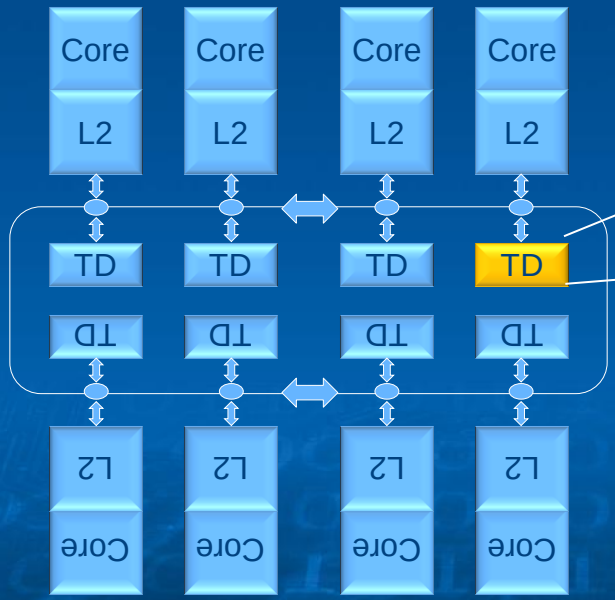
Copyright © 2012 Intel Corporation. All rights reserved.

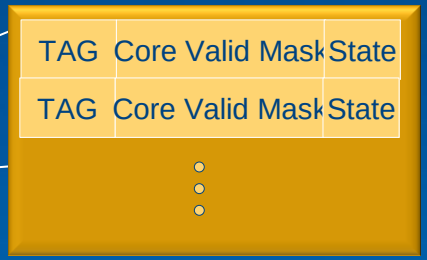
Interconnect



Copyright © 2012 Intel Corporation. All rights reserved.

Distributed Tag Directories

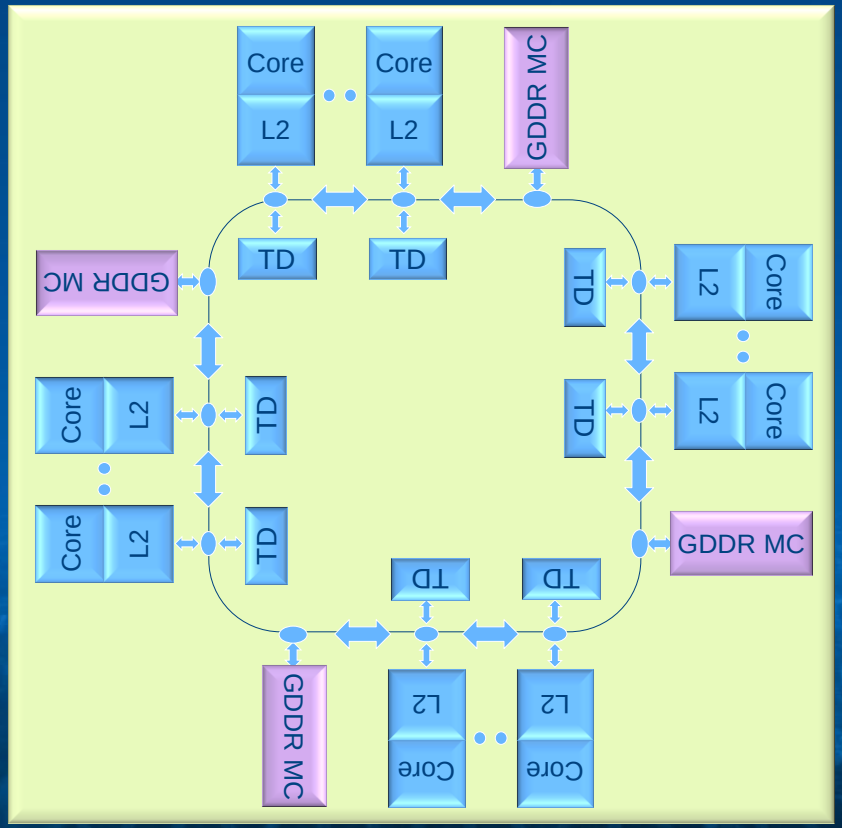




Tag Directories track cache-lines in all L2s

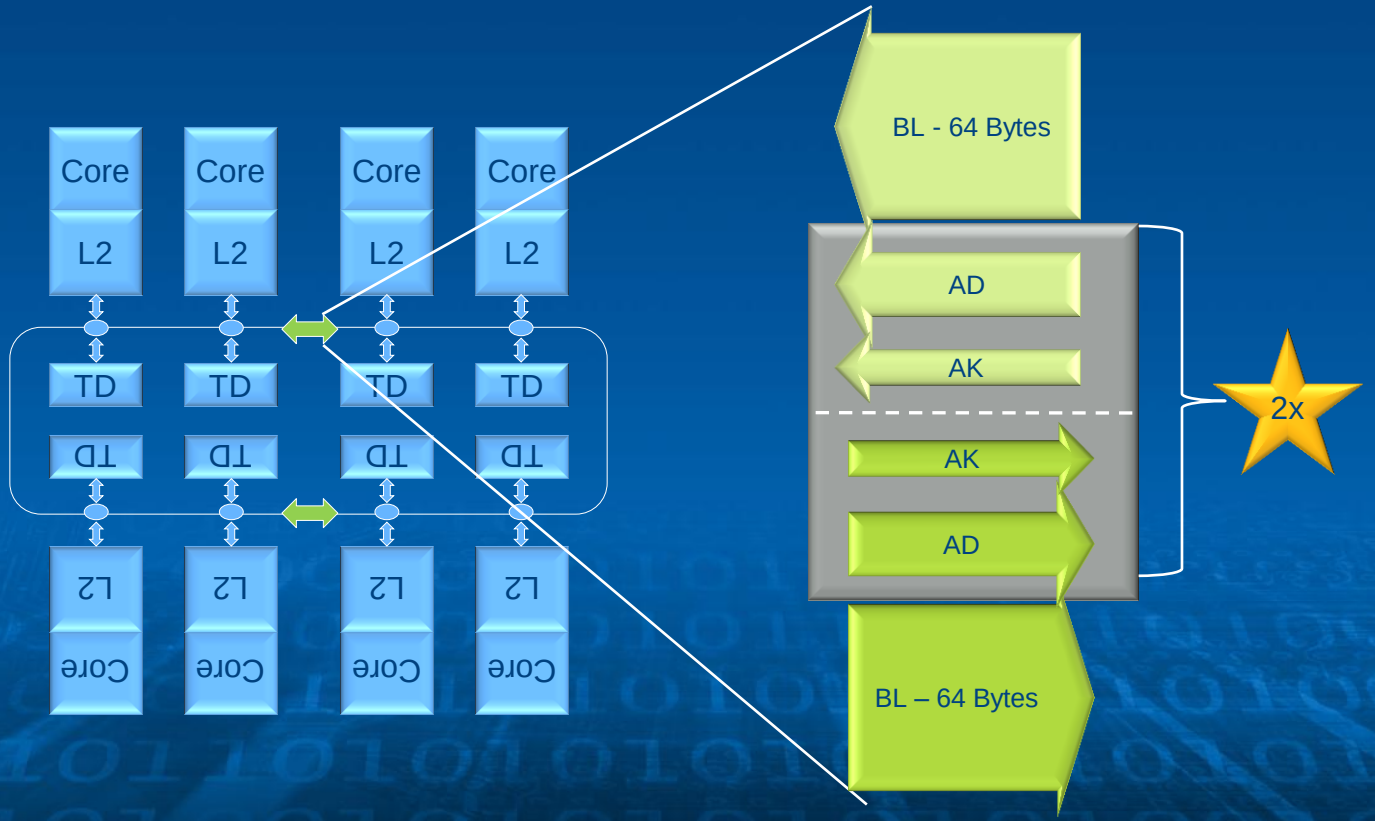
Copyright © 2012 Intel Corporation. All rights reserved.

Interleaved Memory Access



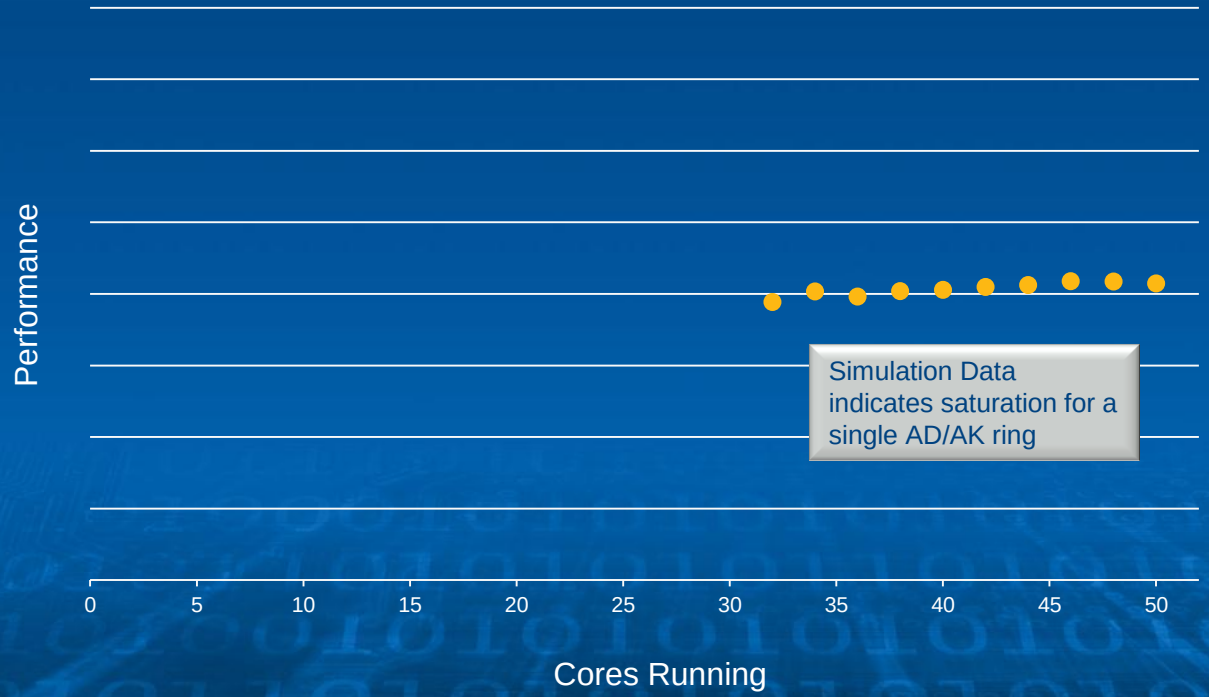
Copyright © 2012 Intel Corporation. All rights reserved.

Interconnect: 2X AD/AK



Copyright © 2012 Intel Corporation. All rights reserved.

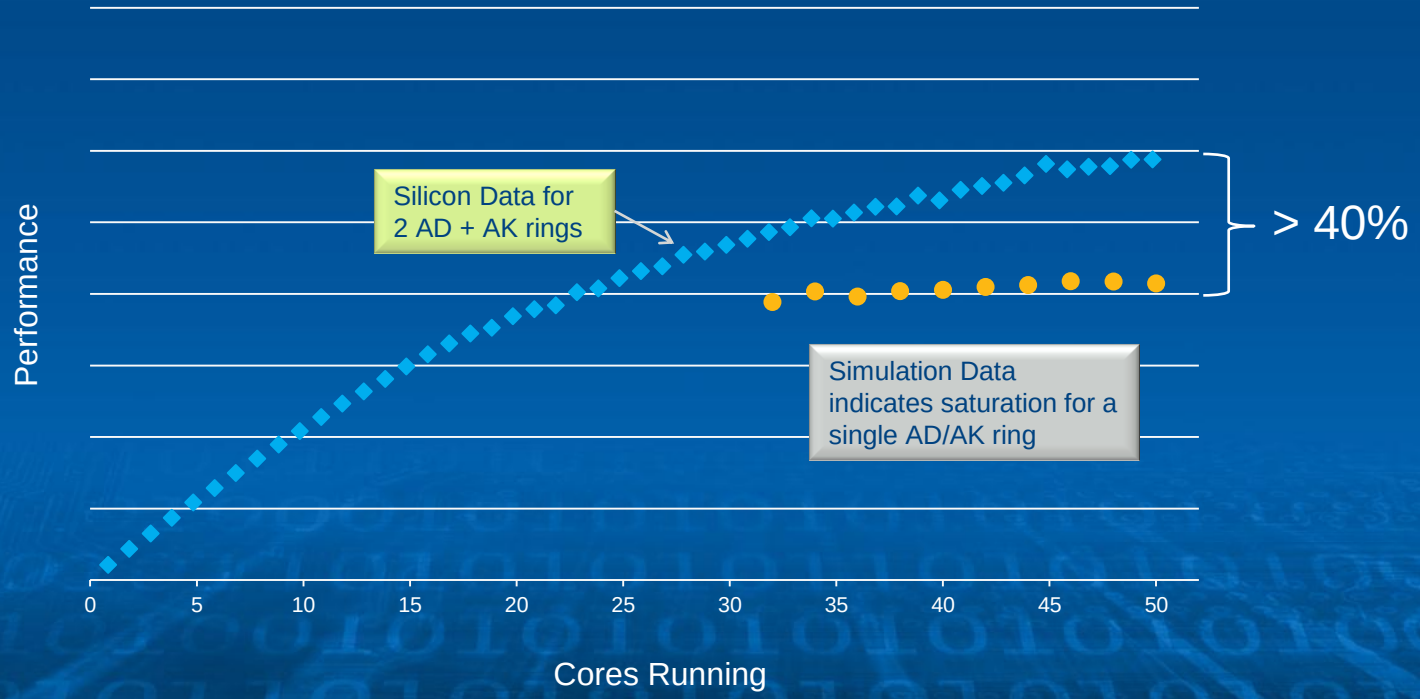
Multi-threaded Triad – Saturation for 1 AD/AK Ring



Results measured in development labs at Intel on Knights Corner prototype hardware and systems. **For more information go to** http://www.intel.com/performance

Copyright © 2012 Intel Corporation. All rights reserved.

Multi-threaded Triad – Benefit of Doubling AD/AK



Results measured in development labs at Intel on Knights Corner prototype hardware and systems. **For more information go to** http://www.intel.com/performance

Copyright © 2012 Intel Corporation. All rights reserved.

Streaming Stores

Streams Triad

```
for (i=0; i<HUGE; i++)
        A[i] = k*B[i] + C[i];
```

Without Streaming Stores

Read A, B, C, Write A

256 Bytes transferred to/from memory per iteration

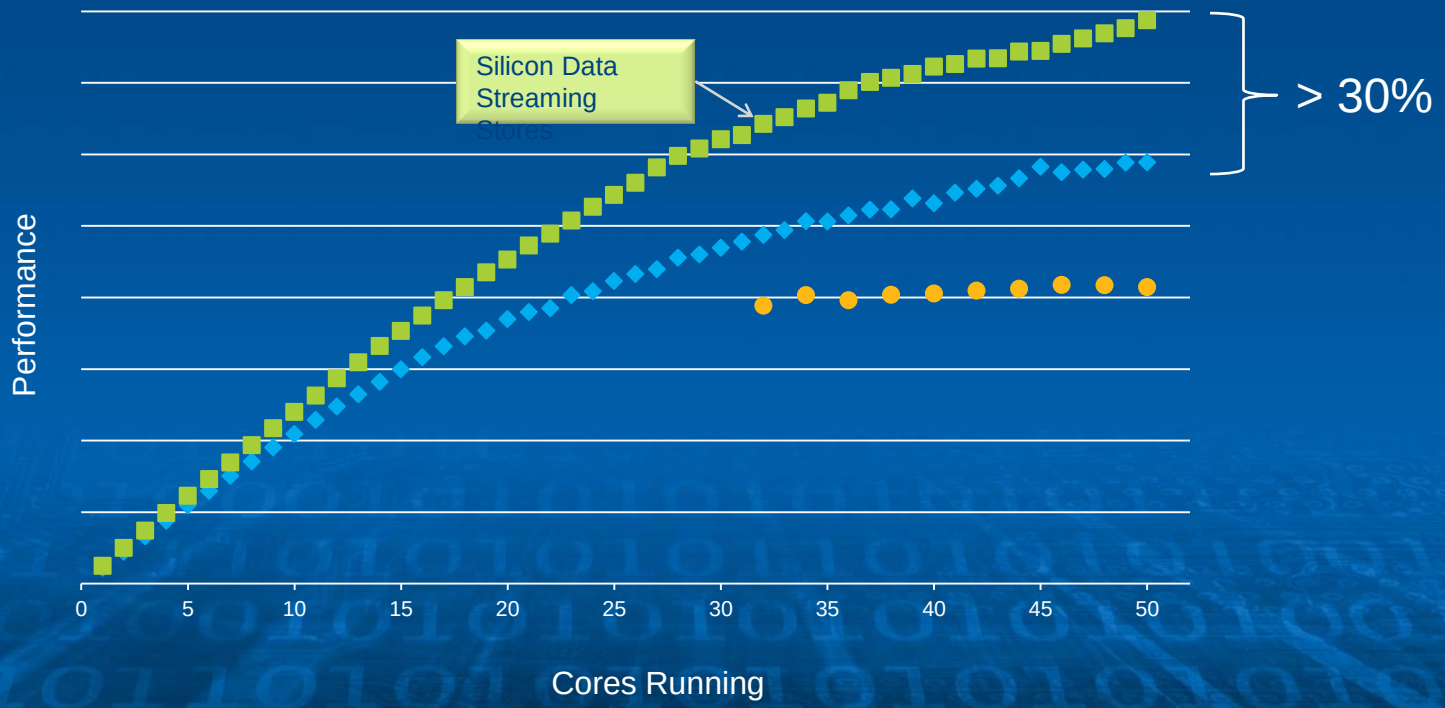
With Streaming Stores

Read B, C, Write A

192 Bytes transferred to/from memory per iteration

Copyright © 2012 Intel Corporation. All rights reserved.

Multi-threaded Triad — with Streaming Stores



Results measured in development labs at Intel on Knights Corner prototype hardware and systems. **For more information go to** http://www.intel.com/performance

Copyright © 2012 Intel Corporation. All rights reserved.

Cache Hierarchy Micro-architecture Choices

L2 TLB
- 64 entry, holds PTEs and PDEs vs. no L2 TLB

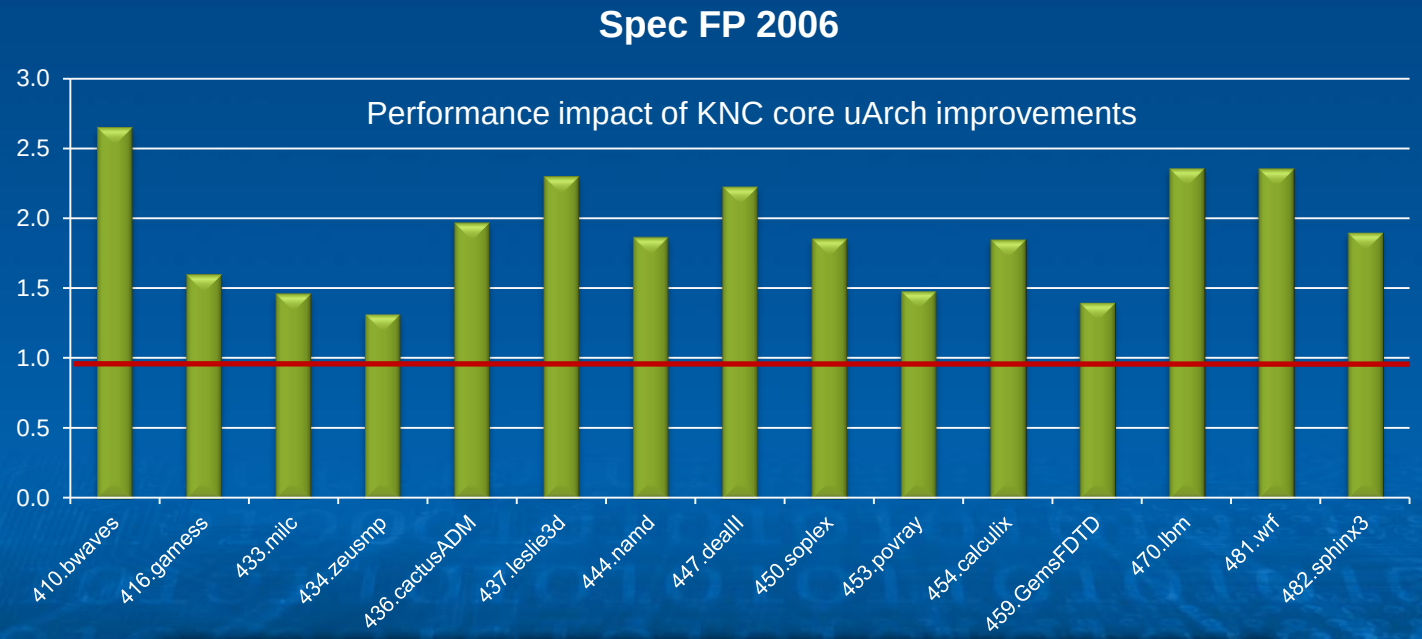
Dcache Capability
- Simultaneous 512b load and 512b store vs. 1 load or store per cycle

L2 Cache
- 512 KB vs. 256 KB

Hardware Prefetcher
- 16 stream detectors, prefetch into the L2 vs. no HWP (rely only on software prefetching)

Copyright © 2012 Intel Corporation. All rights reserved.

Per-Core ST Performance Improvement (per cycle)

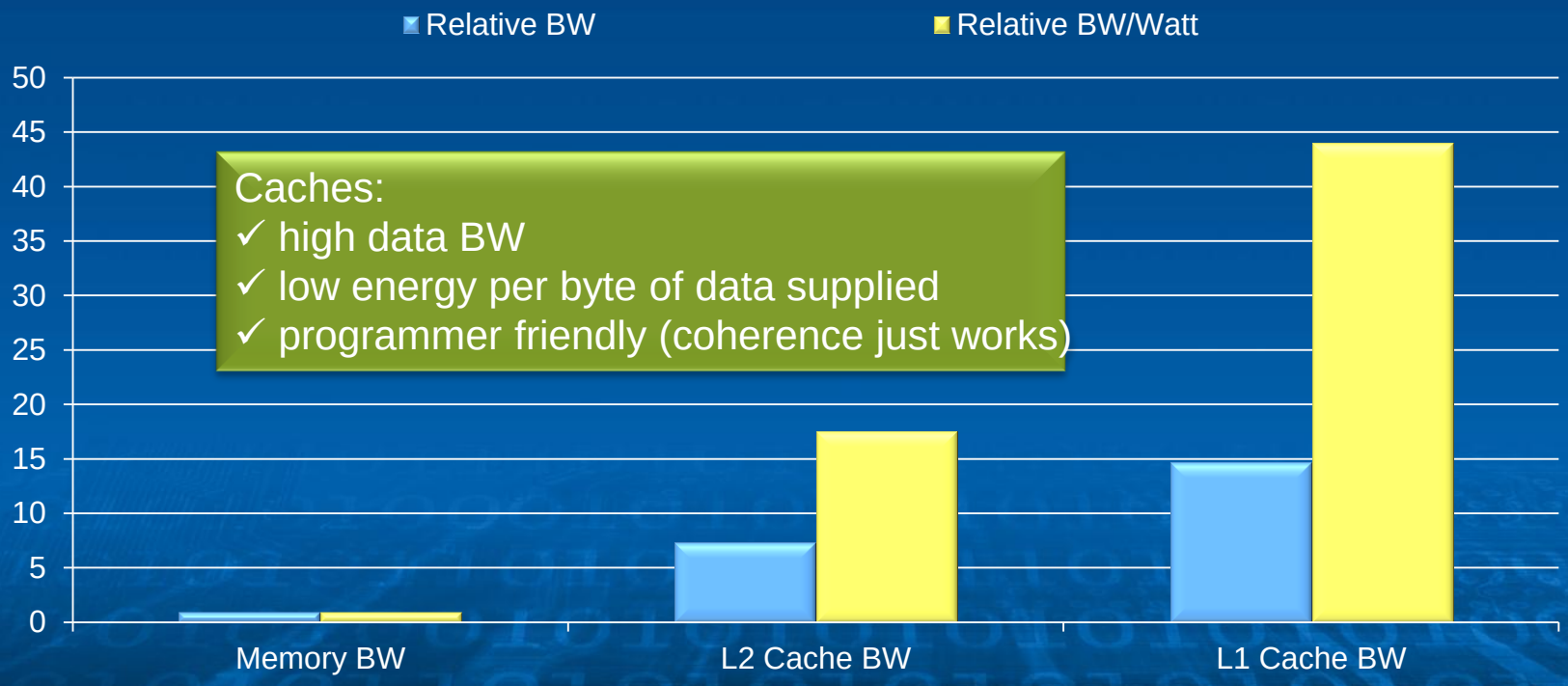


>1.8x Average Performance/Cycle Improvement – 1 Core, 1 Thread

Results measured in development labs at Intel on Knights Corner and Knights Ferry prototype hardware and systems. **For more information go to** http://www.intel.com/performance

Copyright © 2012 Intel Corporation. All rights reserved.

Caches – For or Against?



Caches:
- ✓ high data BW
- ✓ low energy per byte of data supplied
- ✓ programmer friendly (coherence just works)

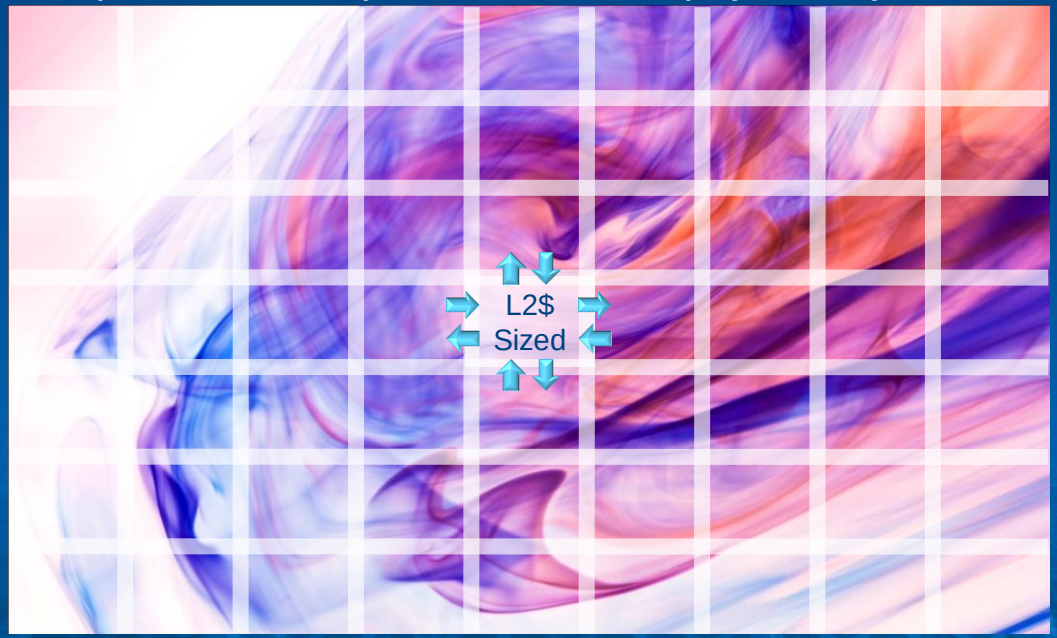
Coherent Caches are a key MIC Architecture Advantage

Results have been simulated and are provided for informational purposes only. Results were derived using simulations run on an architecture simulator or model. Any difference in system hardware or software design or configuration may affect actual performance.

Copyright © 2012 Intel Corporation. All rights reserved.

Example: Stencils

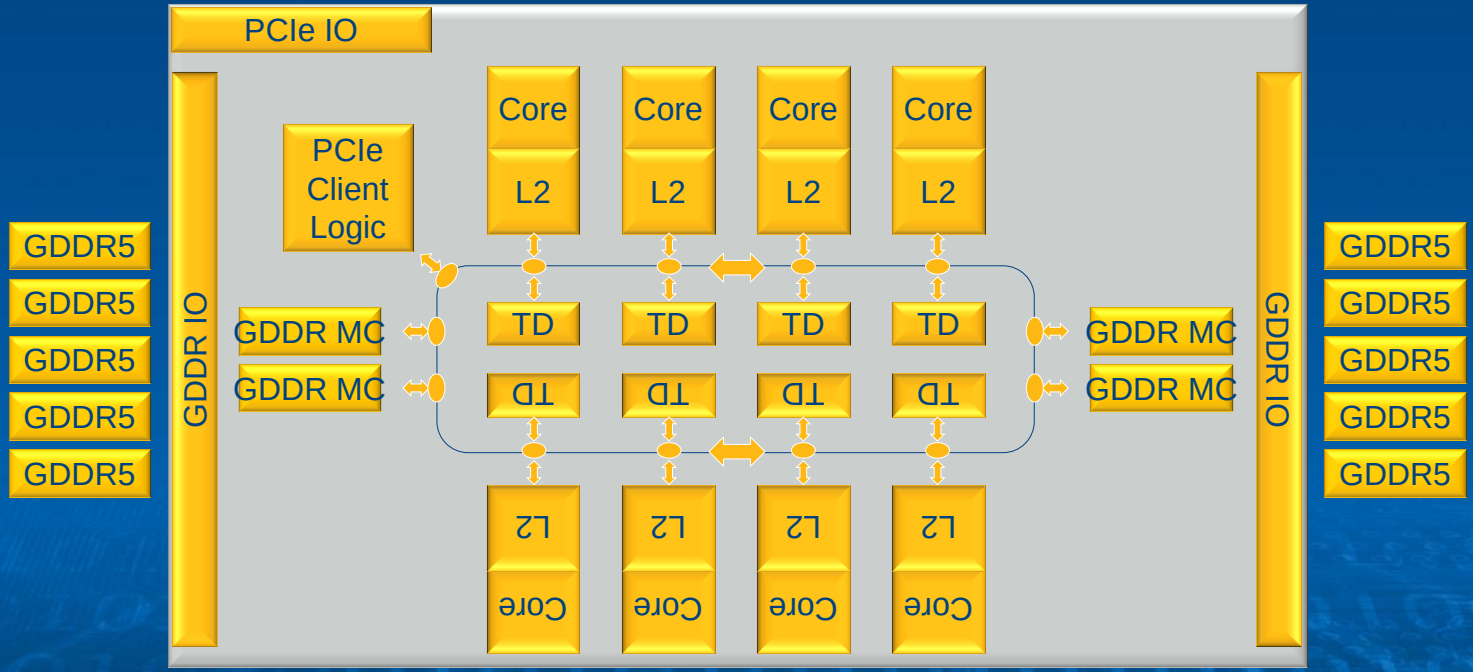
spatial time-step simulation of a physical system



Cache blocking promotes much higher performance and performance/watt vs. memory streaming

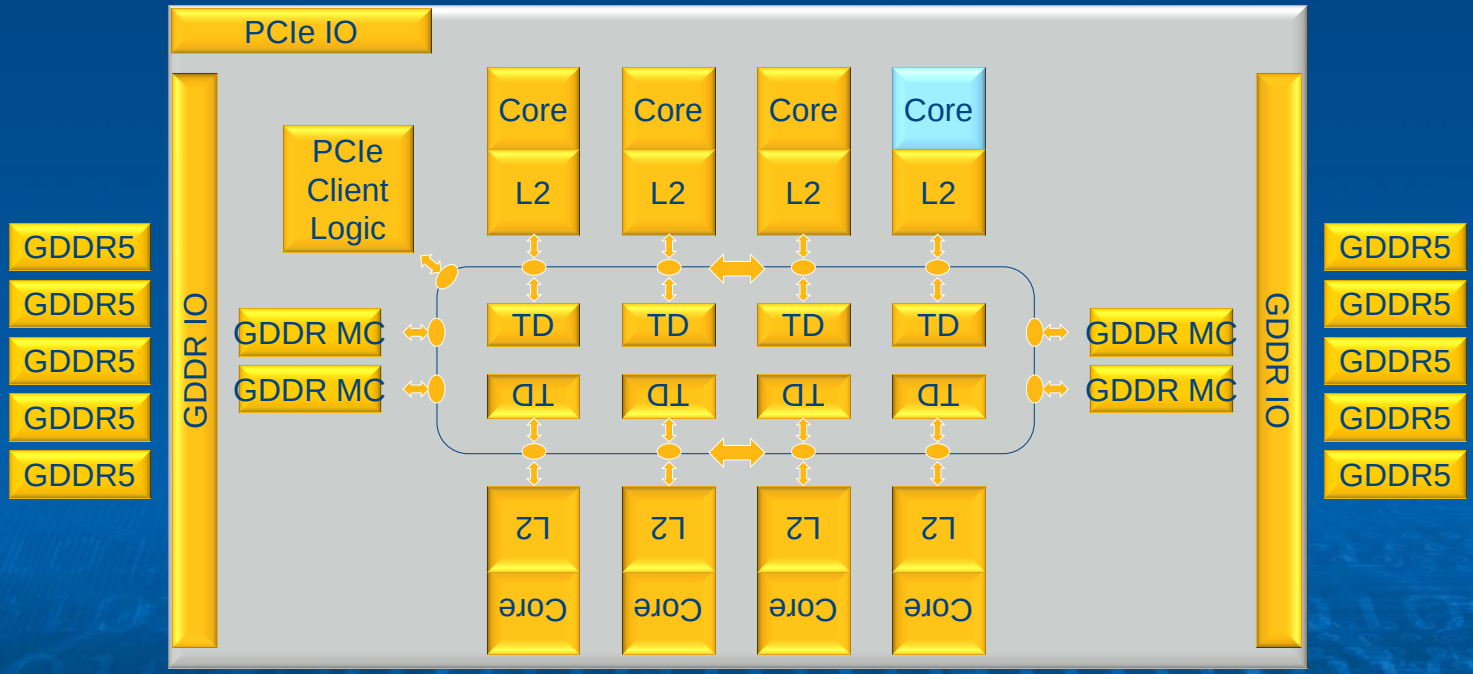
Copyright © 2012 Intel Corporation. All rights reserved.

Power Management: All On and Running



Copyright © 2012 Intel Corporation. All rights reserved.

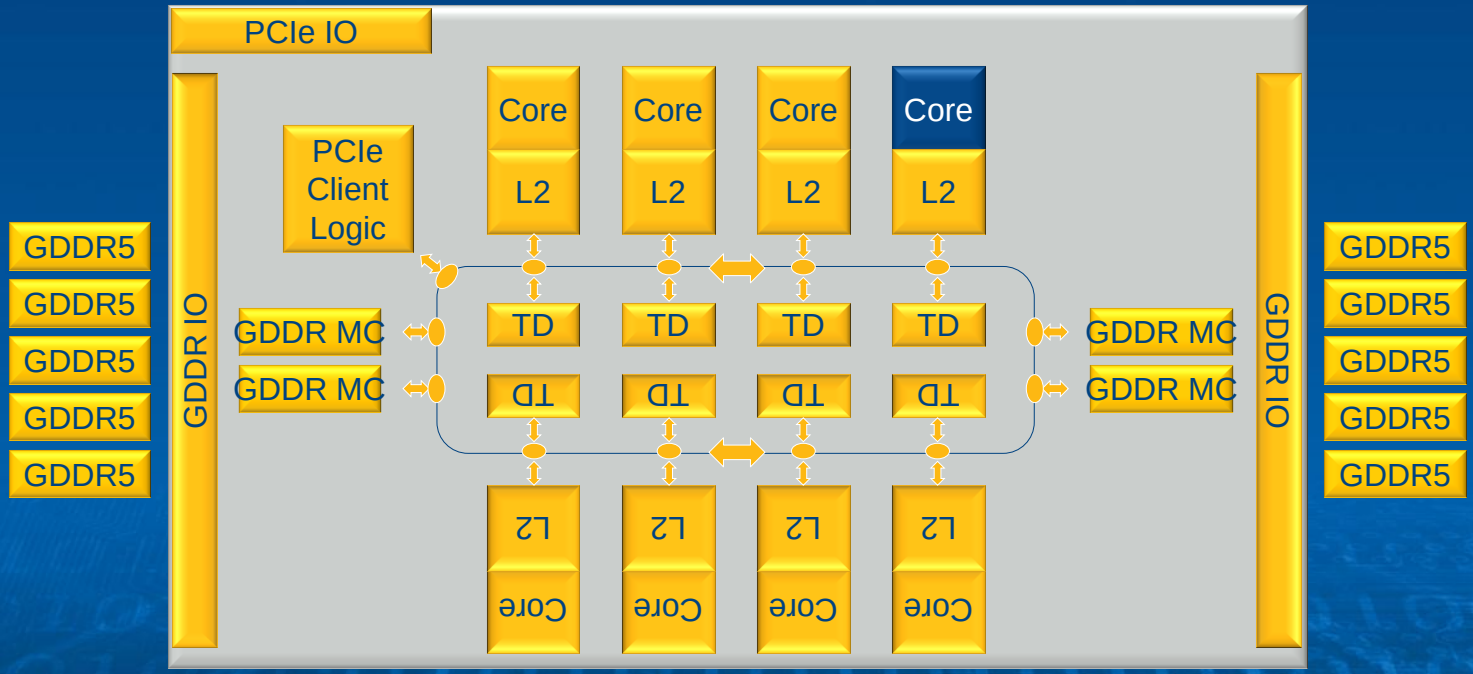
Core C1: Clock Gate Core



When all 4T on a care have halted, core clock gates itself

Copyright © 2012 Intel Corporation. All rights reserved.

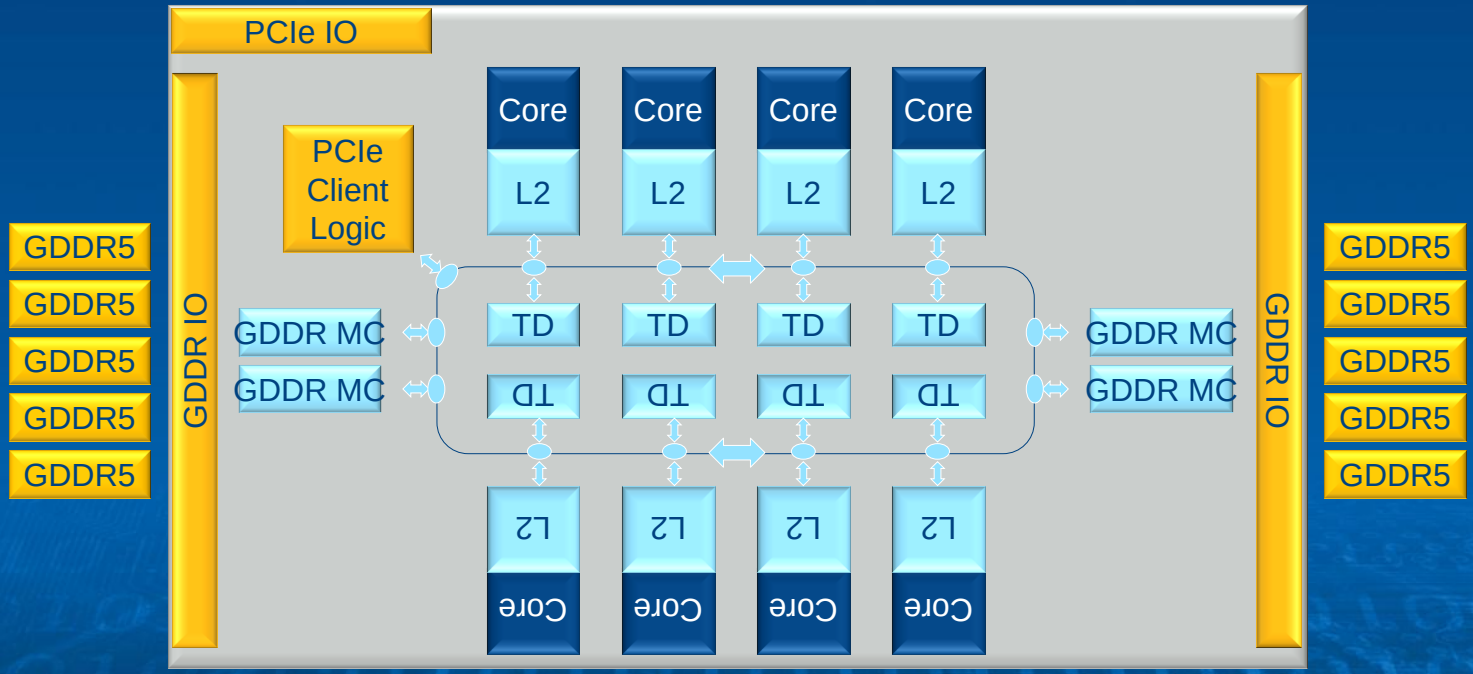
Core C6: Power Gate Core



C1 time-out, power gate core, save leakage, requires core-re-init

Copyright © 2012 Intel Corporation. All rights reserved.

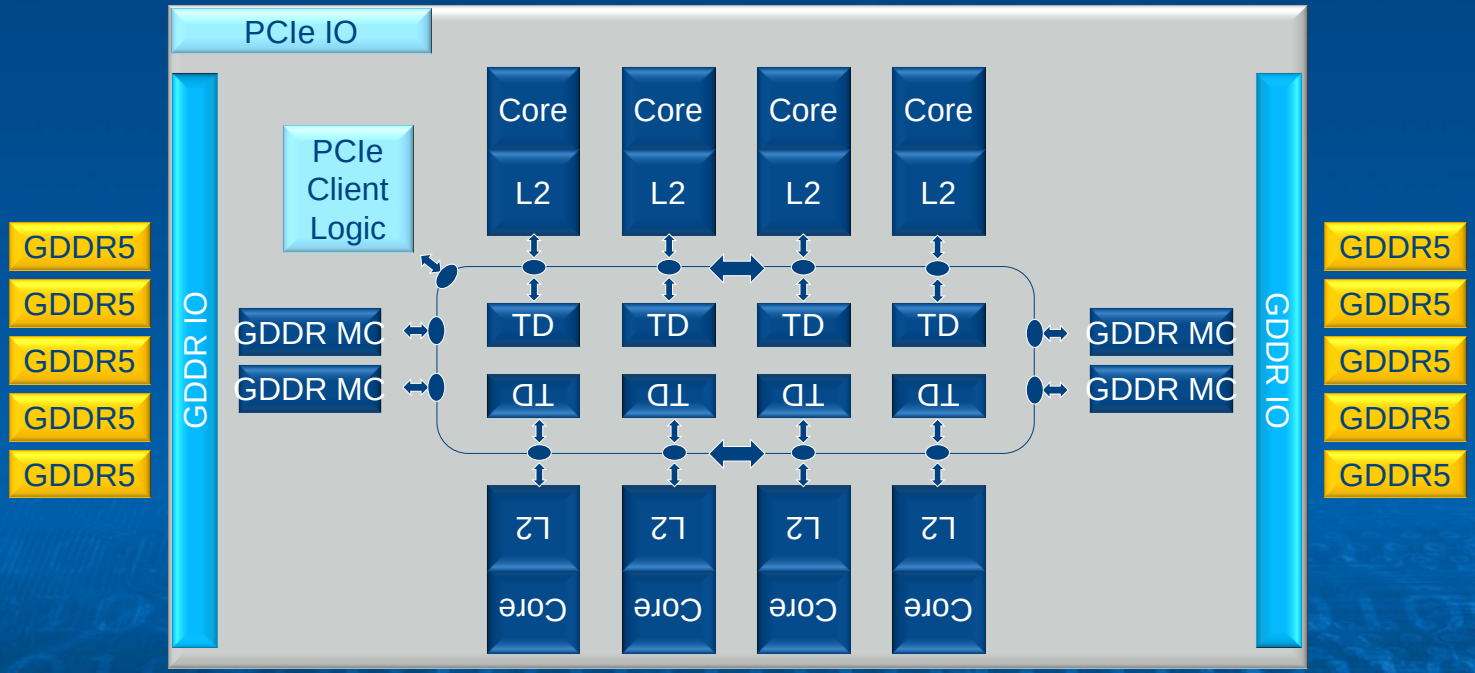
Package Auto C3



Timeout when all cores have been in C6, clock gate the L2 and interconnect

Copyright © 2012 Intel Corporation. All rights reserved.

Package C6



Host Driver can initiate Package C6 –
Uncore Voltage Off, requires partial restart

Copyright © 2012 Intel Corporation. All rights reserved.

Summary

Intel® Xeon Phi™ coprocessor provides:



Performance and Performance/Watt for highly parallel HPC
with cores, threads, wide-SIMD, caches, memory BW

Intel Architecture
general purpose programming environment
advanced power management technology

KNC delivers programmability and performance/watt for highly parallel HPC

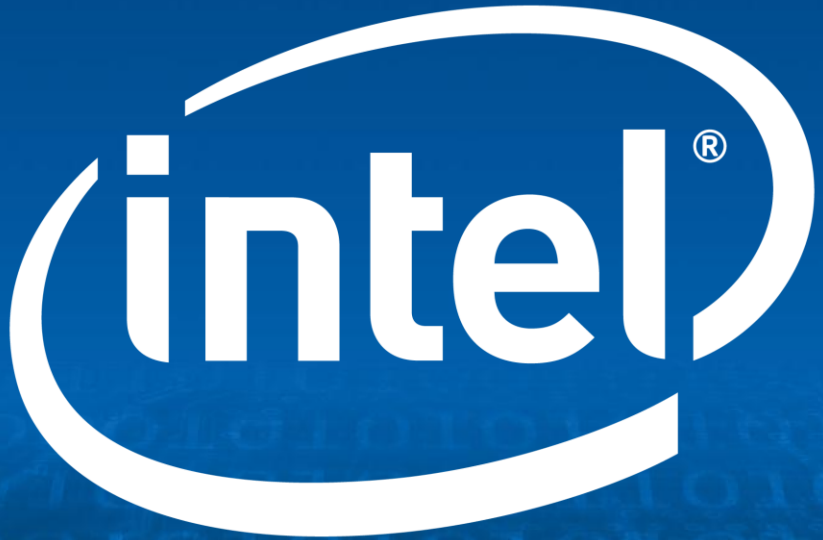
Copyright © 2012 Intel Corporation. All rights reserved.

Thank You

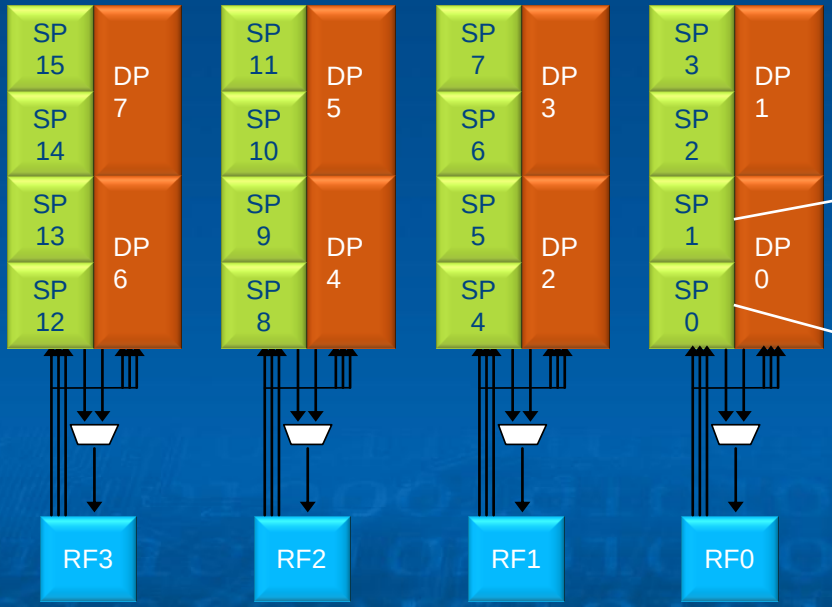
Knights Corner brought to you by:

- IAG (Intel Architecture Group)
 - DCSG (Data Center and Systems Group)
 - VPG (Visual and Parallel Group) MIC
 - HW Architecture
 - HW Design
 - SW
- SSG (Software and Services Group) MIC
- IL PCL (Intel Labs – Parallel Computing Lab)

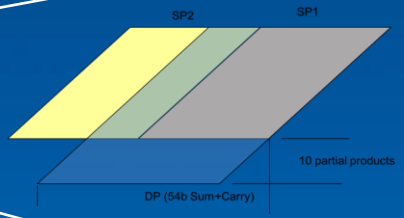
Copyright © 2012 Intel Corporation. All rights reserved.


intel®

Vector Processor: 512b SIMD Width



Shared Multiplier Circuit for SP/DP



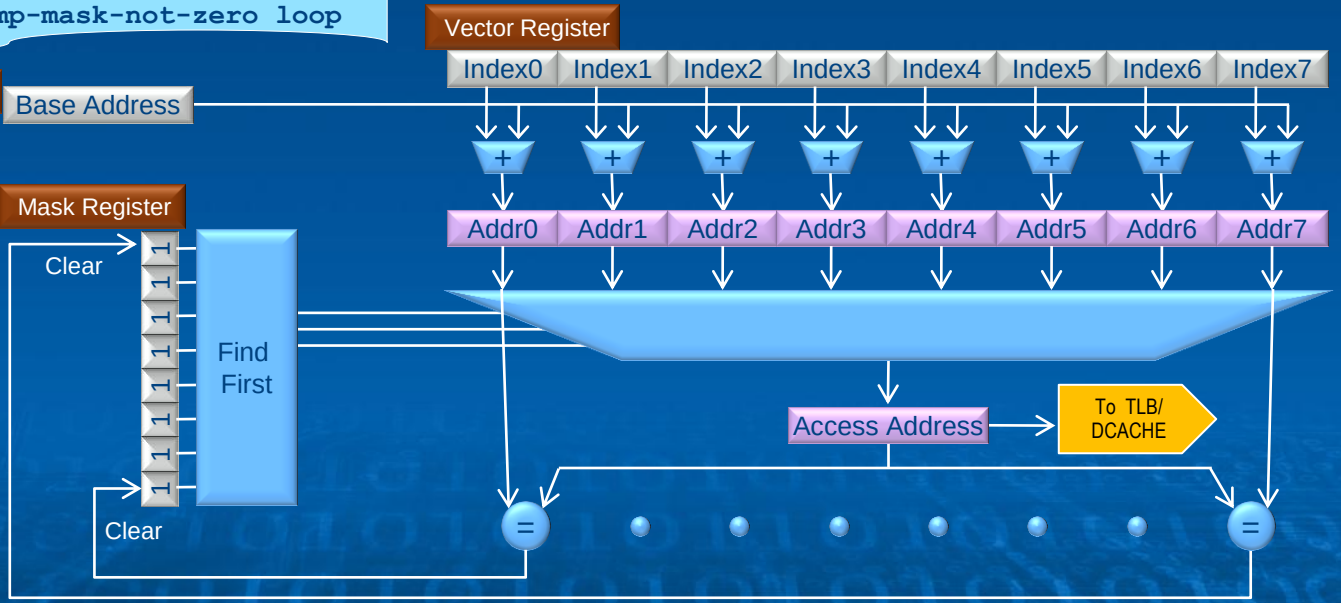
16 wide SP SIMD, 8 wide DP SIMD
2:1 Ratio good for circuit optimization

Copyright © 2012 Intel Corporation. All rights reserved.

Gather/Scatter Address Machinery

Gather Instruction Loop

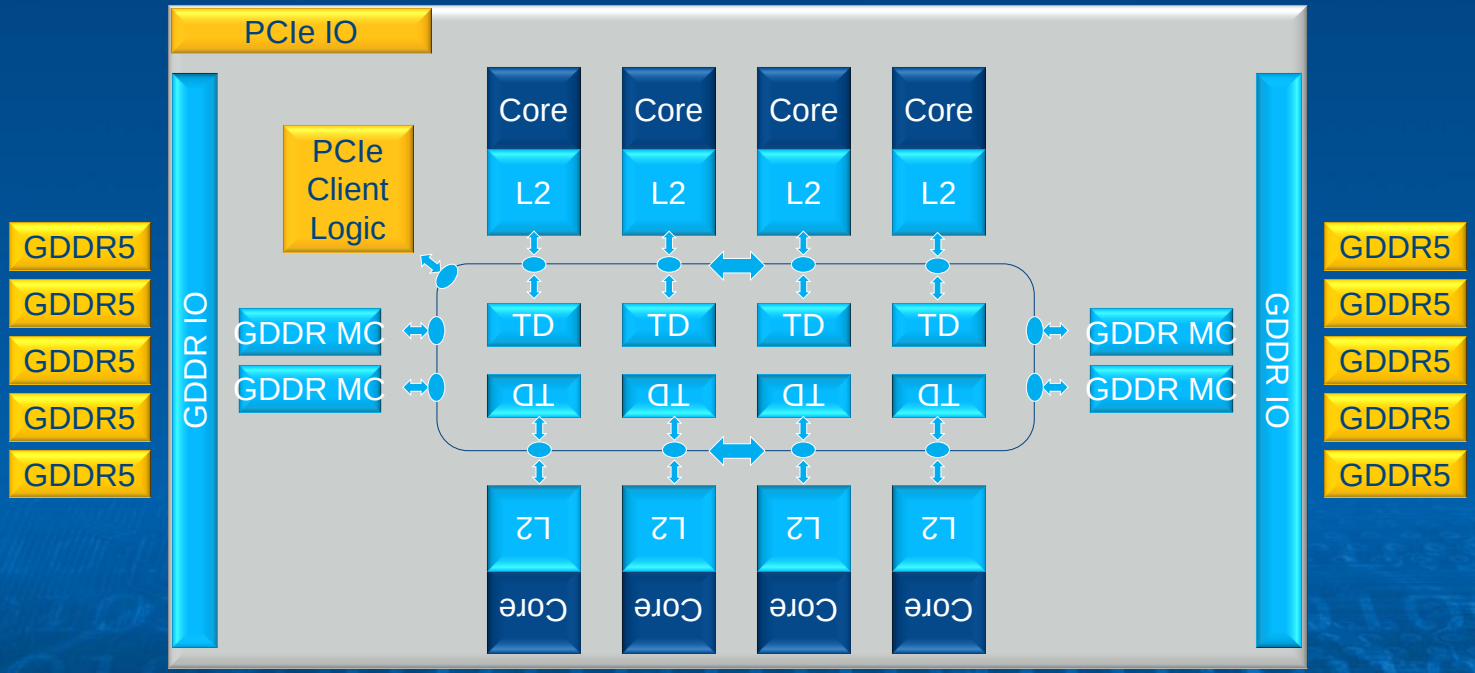
```
     gather-prime
loop: gather-step; jump-mask-not-zero loop
```



Gather/Scatter machine takes advantage of cache-line locality

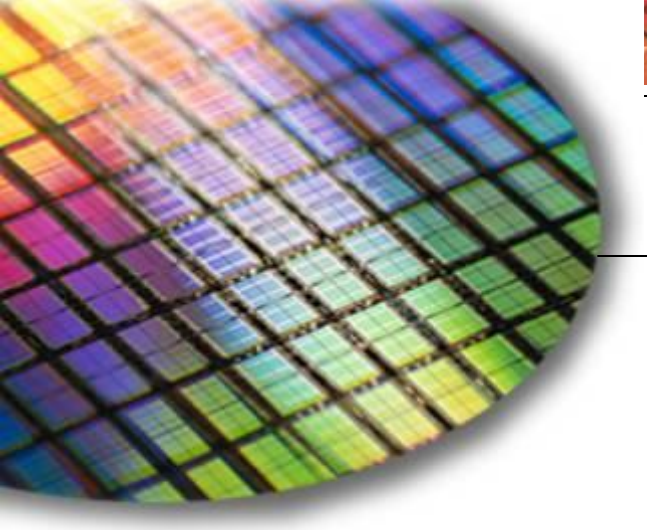
Copyright © 2012 Intel Corporation. All rights reserved.

Package Deep C3



Host Driver Initiated – L2/Ring/TDs dropped to retention V, memory in self refresh

Copyright © 2012 Intel Corporation. All rights reserved.



The World Leader in High Performance Signal Processing Solutions

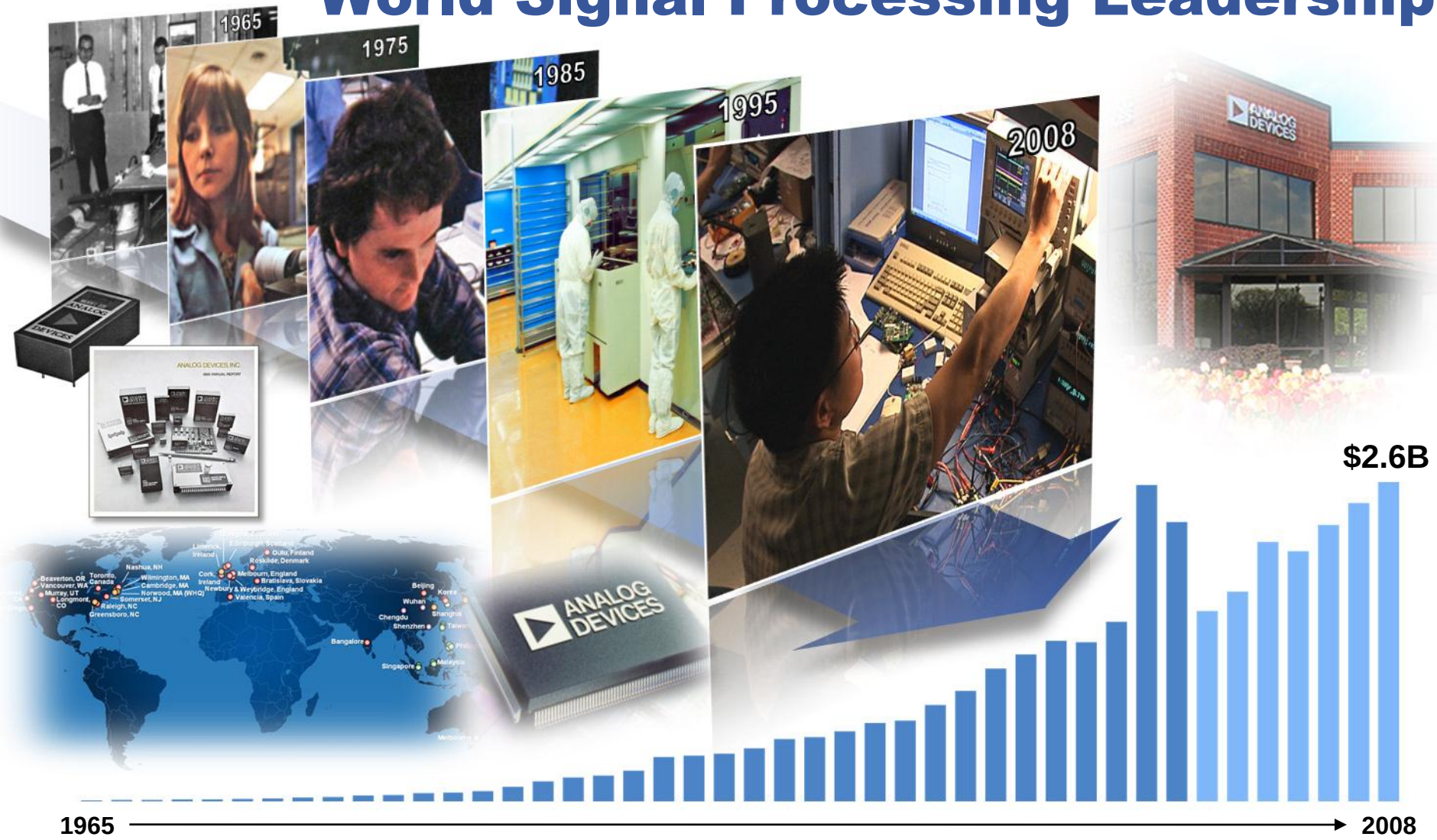
ADI's Revolutionary BF60x Vision Focused Digital Signal Processor System On Chip : 25 Billion Operations/Sec @ 80 mW and Zero Bandwidth

Robert Bushey, Principal Architect & Technologist,

Processor & Digital Signal Processing Core Products & Technologies Group, ADI

ANALOG DEVICES

Innovation Has Driven 40+ Years of Real-World Signal Processing Leadership



Source: *ADI revenue history from ADI financial data. Years 2002-2008 represent continuing operations.*

ANALOG DEVICES

Two New Groups

Collaborating to Address Customer Needs in Market-specific Ways

Core Products & Technologies Group

Converters



MEMS/ Sensors



Processor-DSP



Power



Linear & RF





Share Segment and Applications Expertise to Enhance Core Technologies

Integrate Core Technologies into Solutions for Key Segments and Applications

Strategic Market Segments Group



Automotive



Communications Infrastructure



Consumer



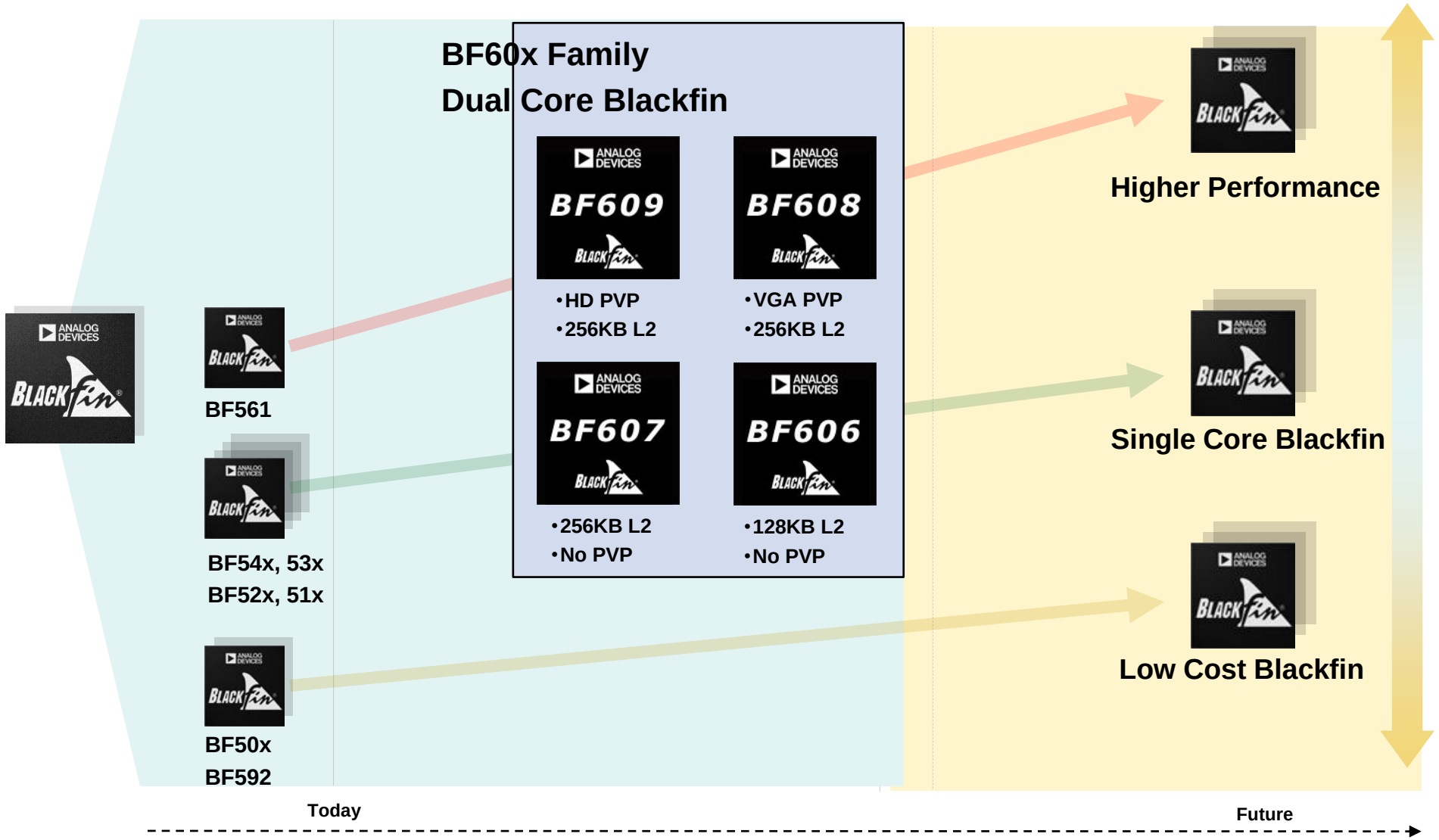
Healthcare



Industrial & Instrumentation

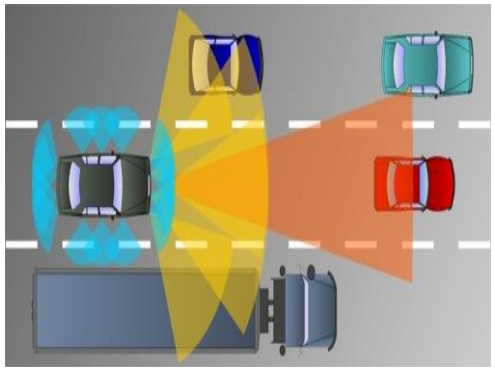
ANALOG DEVICES

Analog Devices Blackfin Processor Roadmap



Advanced Driver Assistance Systems(ADAS)

- RAdio Detection And Ranging (RADAR)
 - Object detection system using electromagnetic waves to calculate range, height, direction and/or speed of fixed and moving objects.
- LIght Detection And Ranging (LIDAR)
 - Optical sensing technology that measures properties of scattered light to find range and/or other information about a distant object.



- Vision Processing / Video Recognition
 - Requires a Very High Performance Real Time Digital Signal Processing Solution
 - ✓ Pre-crash Warning and/or Avoidance
 - ✓ Lane Departure Warning (LDW)
 - ✓ Traffic Sign Recognition (TSR)
 - ✓ General object classification, tracking & verification



 - ➔ Customer & Market Driven DSP Requirements
 - ✓ Real Time @ 30FPS at 1280x960 Pixels/Frame Performance
 - ➔ 37 Megapixels / Second Real Time ADAS Analytics
 - ➔ Many Parallel and Serial Concurrent Operations / Pixel
 - ➔ BILLIONS of Operations / Sec or GOPS
 - ✓ Low Power, Low Cost, and Low Bandwidth Constraints

ANALOG DEVICES

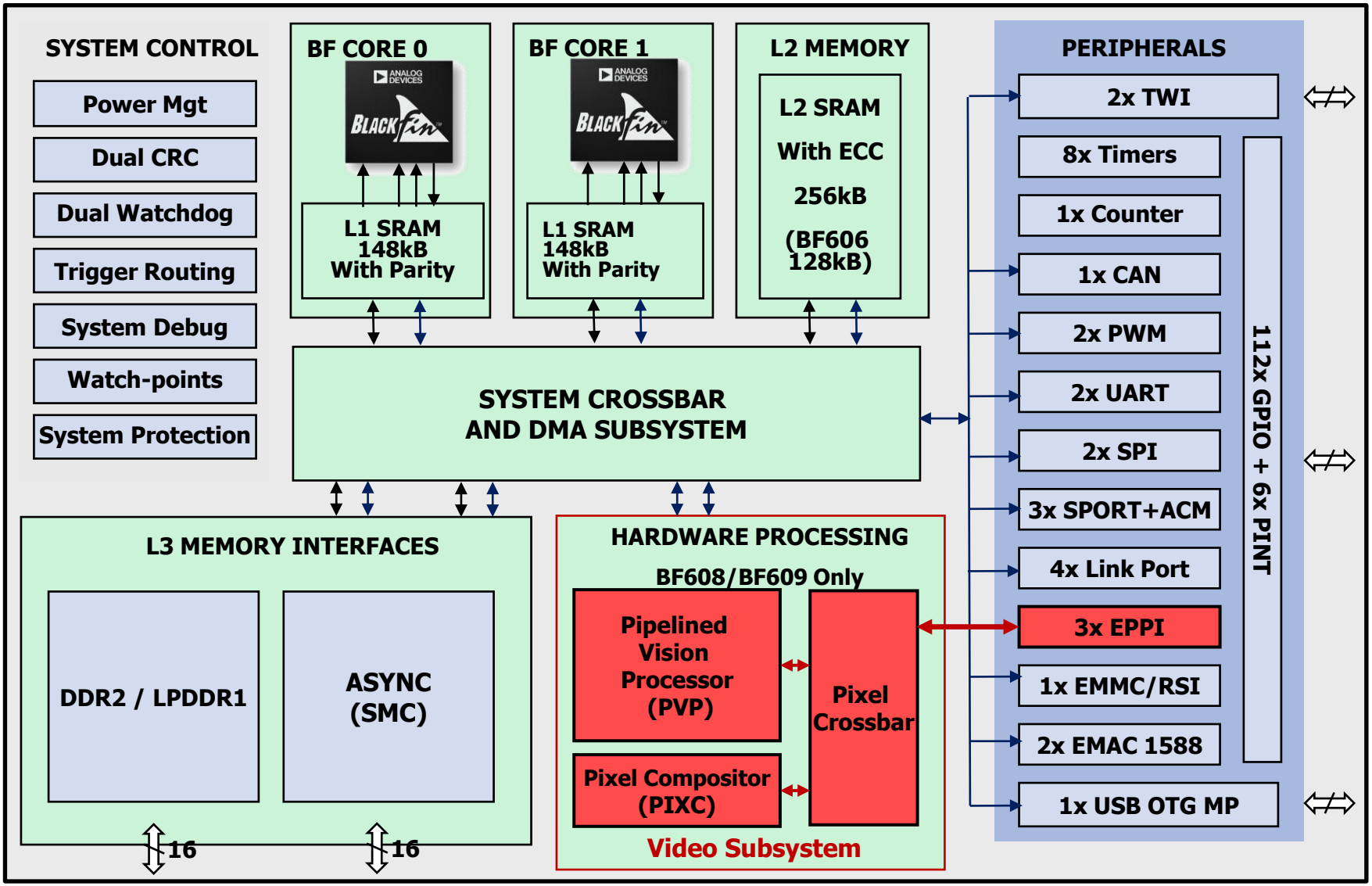
ADSP-BF609 Blackfin Highlights (1)

- **New Pipelined function-level Vision Processor (PVP) for embedded vision applications**
 - **Supports multiple concurrent analytics functions at low price with low power consumption**
 - With our new dedicated function level vision processor, broad adoption of sophisticated, multi-function analytics can now be feasibly deployed into all levels of embedded vision applications
- **Highest performance Blackfin Instruction-level processing**
 - 1GHz of programmable Blackfin instruction level processor performance across two cores
 - Large on-chip memory : 4.3Mbit SRAM & highly efficient system bandwidth

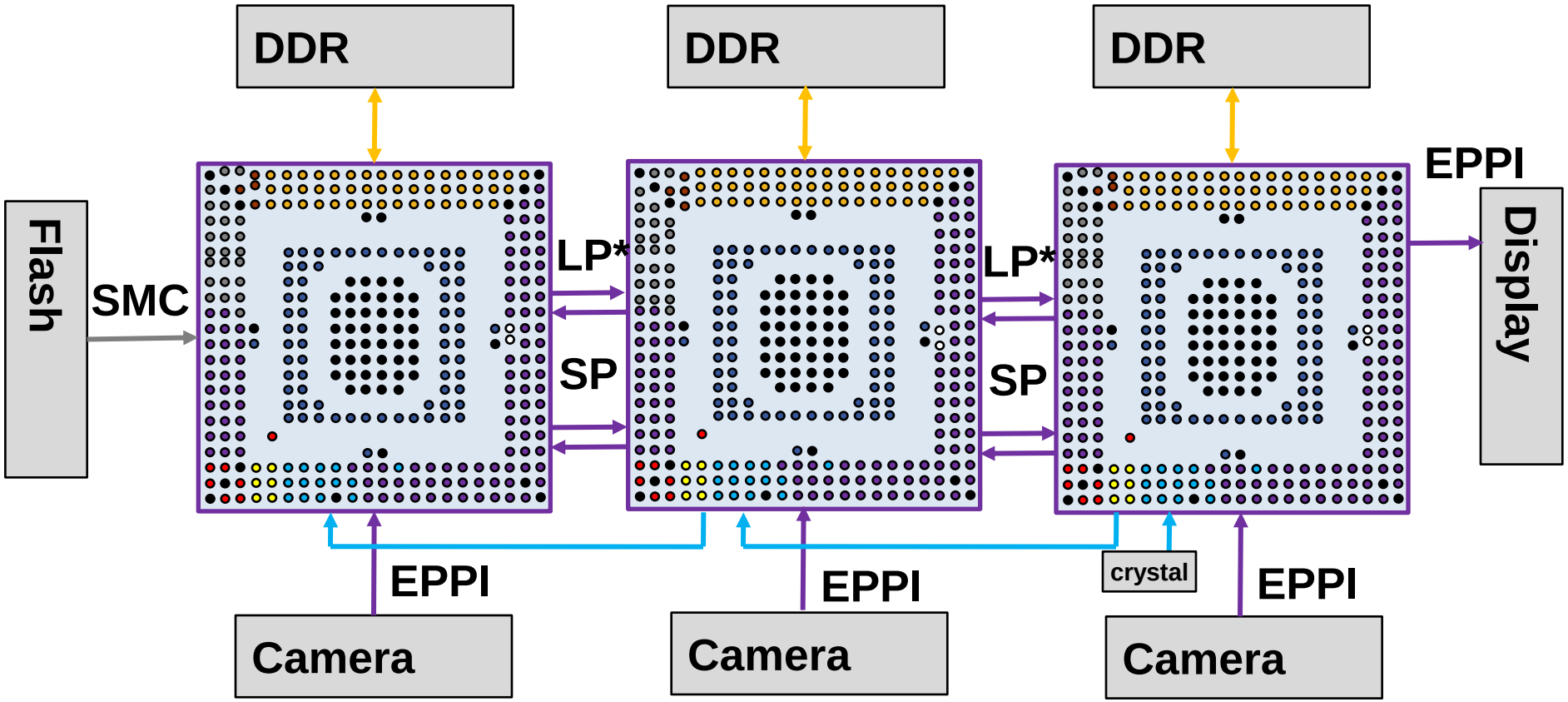
ADSP-BF609 Blackfin Highlights (2)

- **Feature rich peripheral set & connectivity options**
 - Memory interfaces: DDR2, LPDDR, RSI **(Removable Storage Interface for MMC, SD, SDIO, and CE-ATA)**
 - Connectivity: USB2.0, Ethernet, 5 types of serial interfaces, ePPI Video Interface for seamless CMOS sensors and LCD connectivity and control
 - Link ports for high speed multiprocessing and inter-chip communication
- **Integration for safety oriented applications**
 - Memory parity, ECC, system protection unit for detecting/recovering from faults
- **Delivering lowest power per function**
 - Typical power consumption at 25C for the BF609 is 400mW

BF609 Block Diagram



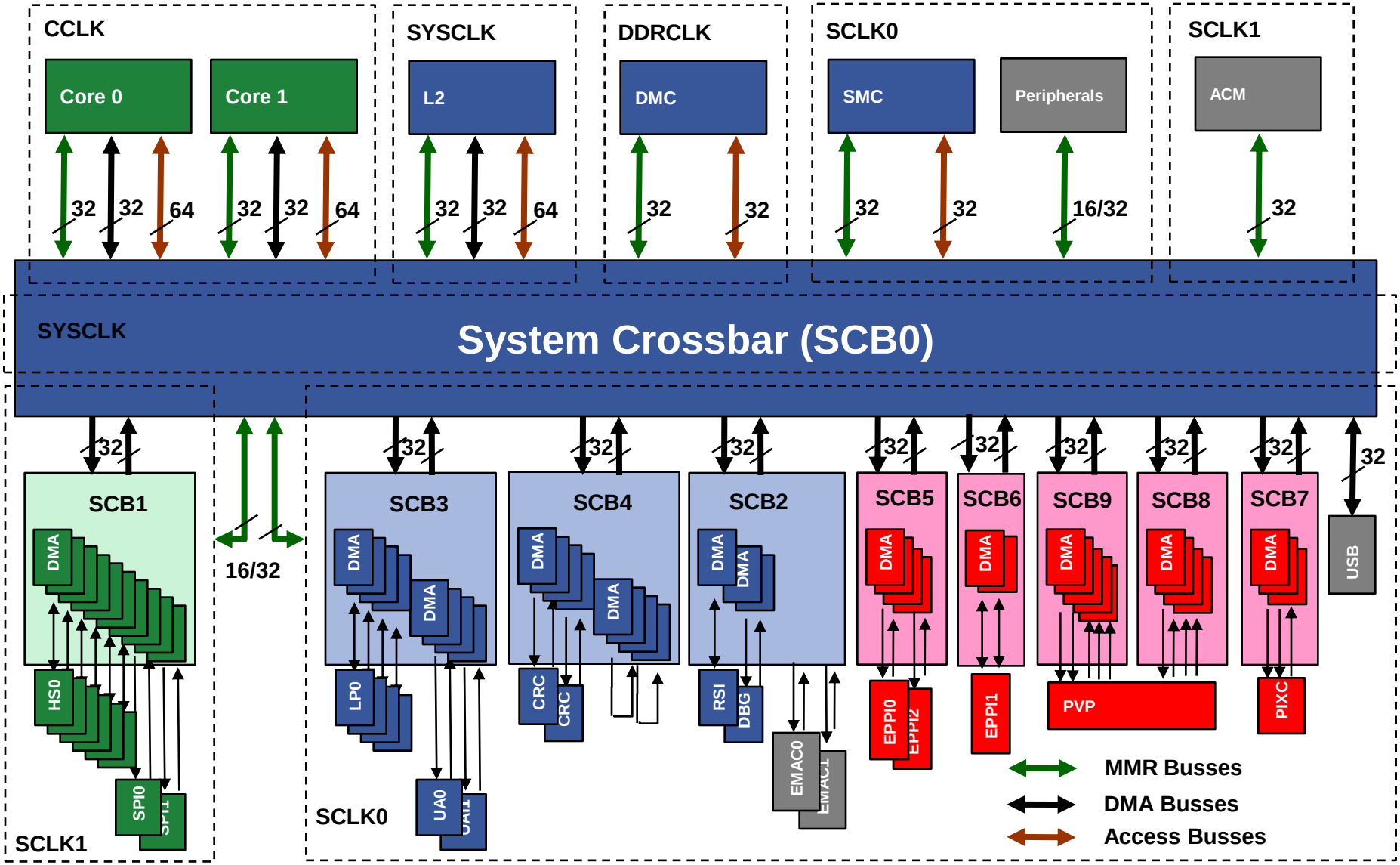
BF609 is Optimized for Many-way Multi-processing With Efficient Inter- chip Communication and Control



* or unidirectional 16-bit PPI

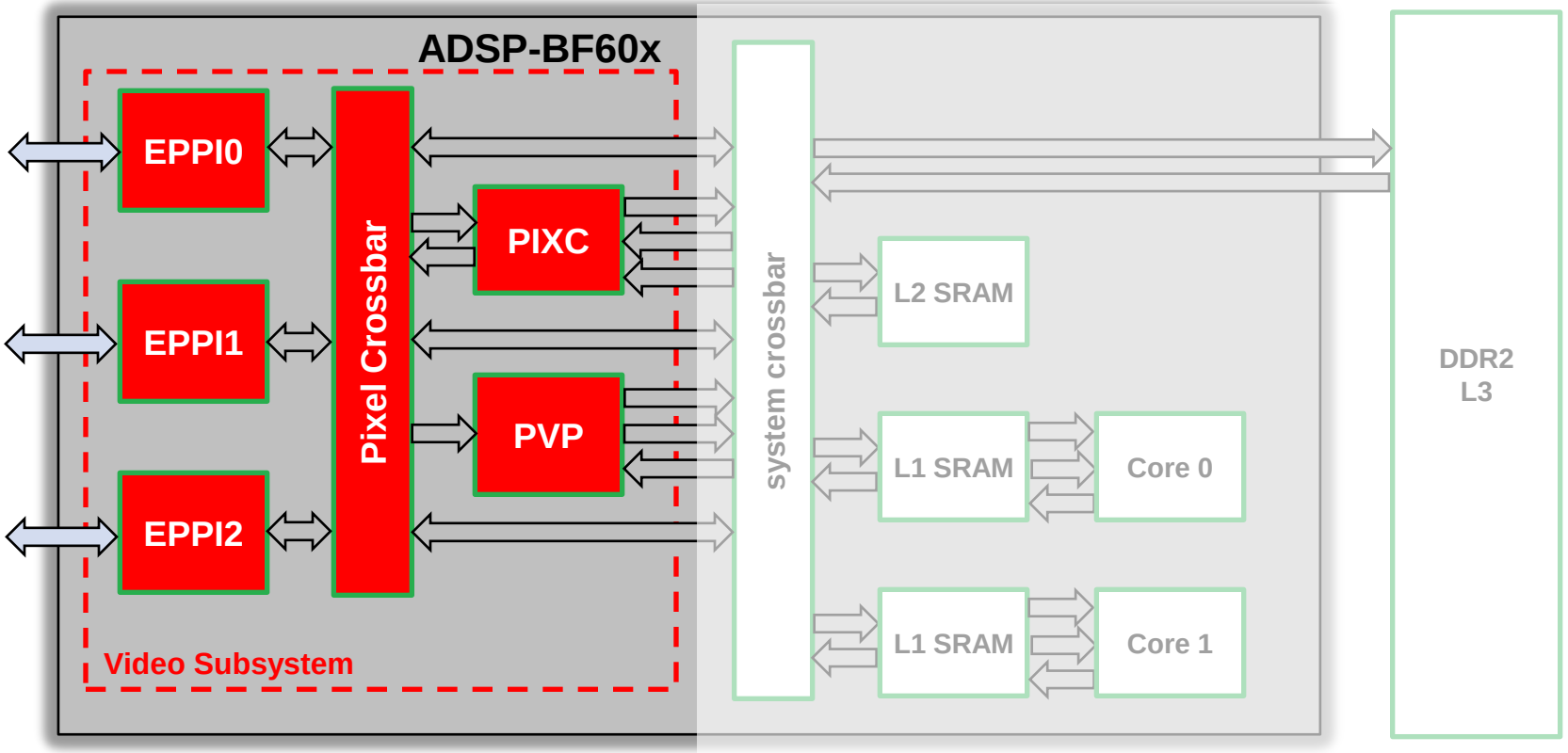
ANALOG DEVICES

BF609's Masters, Slaves, And Interconnect



ANALOG DEVICES

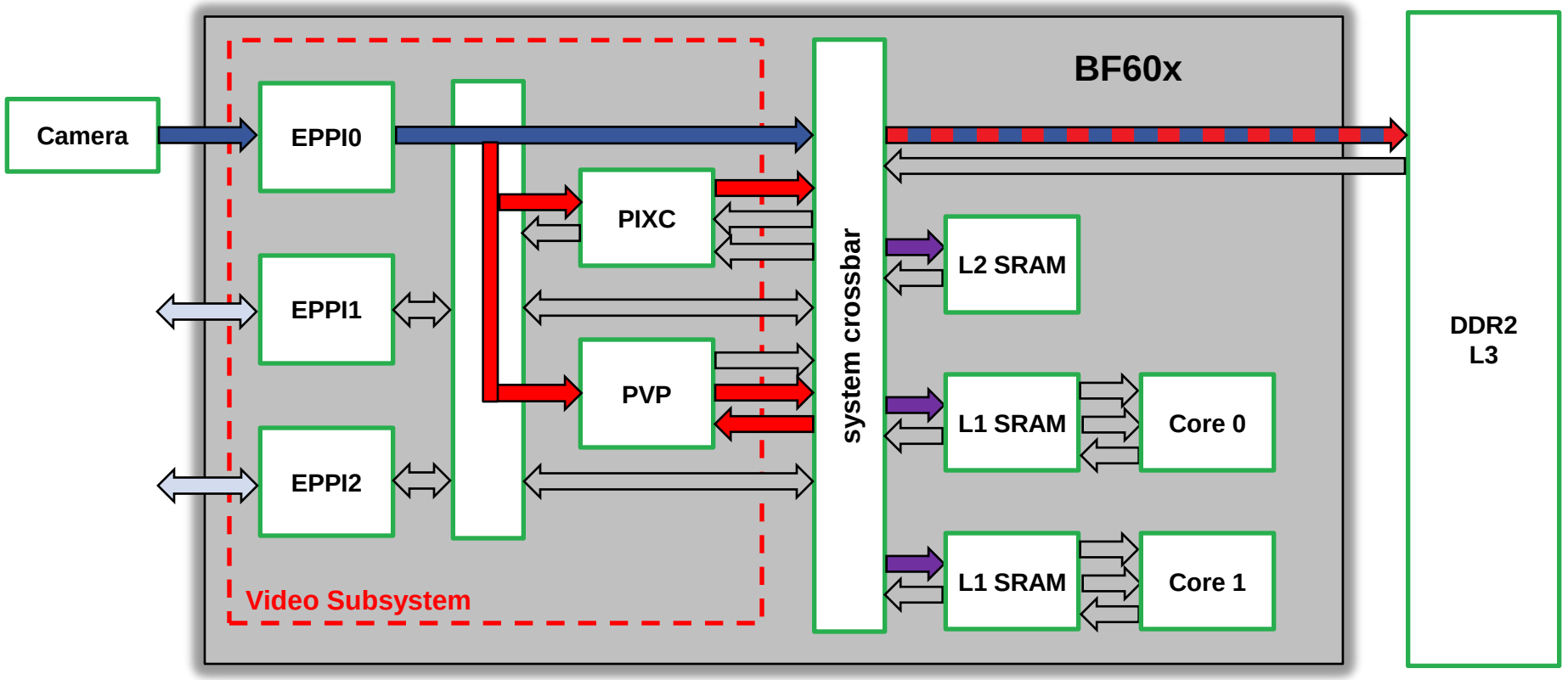
BF609 Video Subsystem & Interconnect



- BF60x Introduces a new Video Subsystem(VSS) architecture and interconnect:
 - 3 Enhanced Parallel Peripheral Interfaces (EPPI);
 - Pipelined Vision Processor (PVP);
 - Pixel Compositor (PIXC);
 - Pixel Crossbar

ANALOG DEVICES

Video Subsystem Vision Processing Datapath



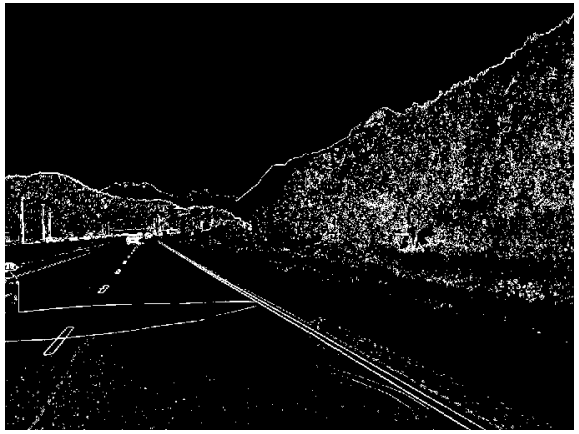
- Data is processed / analyzed before it goes to memory
 - Traffic does not load system bandwidth / power / EMI
- Raw PPI data can go to memories in parallel
 - Data broadcasted to DMA, PIXC and PVP;
 - Multiple data pathes distribute to L1 / L2 / L3 memories

ANALOG DEVICES

Pipelined Vision Processor(PVP) Overview

- The PVP computes more than 25 billion operations per second of vision processing while consuming little power and utilizing limited memory bandwidth
 - Used in Advanced Driver Assistance, Robotic and Machine Vision Systems, as well as other adjacent vision/imaging applications
- PVP provides application performance across the following major areas:
 - Object Detection
 - Object Classification and Tracking
 - Object Verification
- PVP works in conjunction with the high performance instruction level programmable Blackfin DSP cores
- PVP reduces required off-chip bandwidth by windowing and pre filtering input data



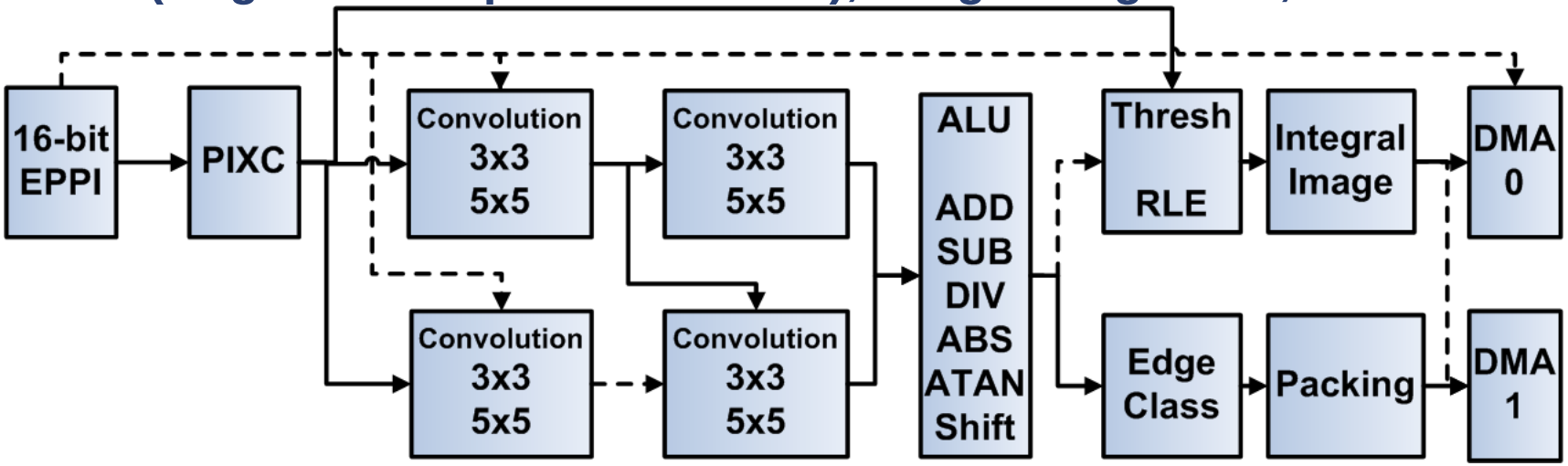


Example
Canny or Sobel Edge Detection

ANALOG DEVICES

Pipelined Vision Processor(PVP) Key Features, Pixel Data Path Flexibility & Function Level Processing Capabilities

- Optimal bandwidth reduced pixel datapaths
- Function level processing with highly configurable datapath
- Enables many computationally complex vision applications
- Allows for concurrent support of multiple applications
- Supported Image size (frame rate 30 fps): 1280x960, 1024x768, 640x480
- Supported Pixel-width: up to 16bits
- PVP Supports Vision Function Level Processing: Sobel filter & Canny filter(Convolution), Histogram, ARCTAN and Absolute Value(Angle and amplitude vectors), Image integration, Pixel



ANALOG DEVICES

Pipelined Vision Processor (PVP) Key Function Level Processing Blocks (1)

- **2D Convolution Blocks**
 - Supports 1*1, 3*3, 5*5 configurations up to 16-bit input, 16-bit coefficient (updateable line by line)
 - Internal 37-bit Acc & Barrel shift
 - Scaled to 32-bit result
 - PVP Initialization via zero filled lines or duplication of the first/last line per frame

- **ALU/Cartesian to Polar Block**
 - Input two data-streams at 32-bit
 - Output two 16-bit streams or one 32-bit stream
 - Math operations supported (signed/unsigned)
 - ADD, SUB, 32-bit multiply, 32-bit divide, Accumulation (xx bit)
 - Shift (logic, arithmetic), XOR, Masking, Inversion, Arctan, Absolute value (x2+y2)

ANALOG DEVICES

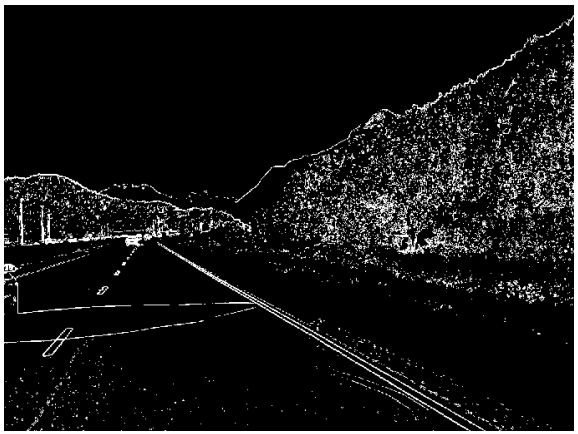
Pipelined Vision Processor (PVP)
Key Function Level Processing Blocks (2)

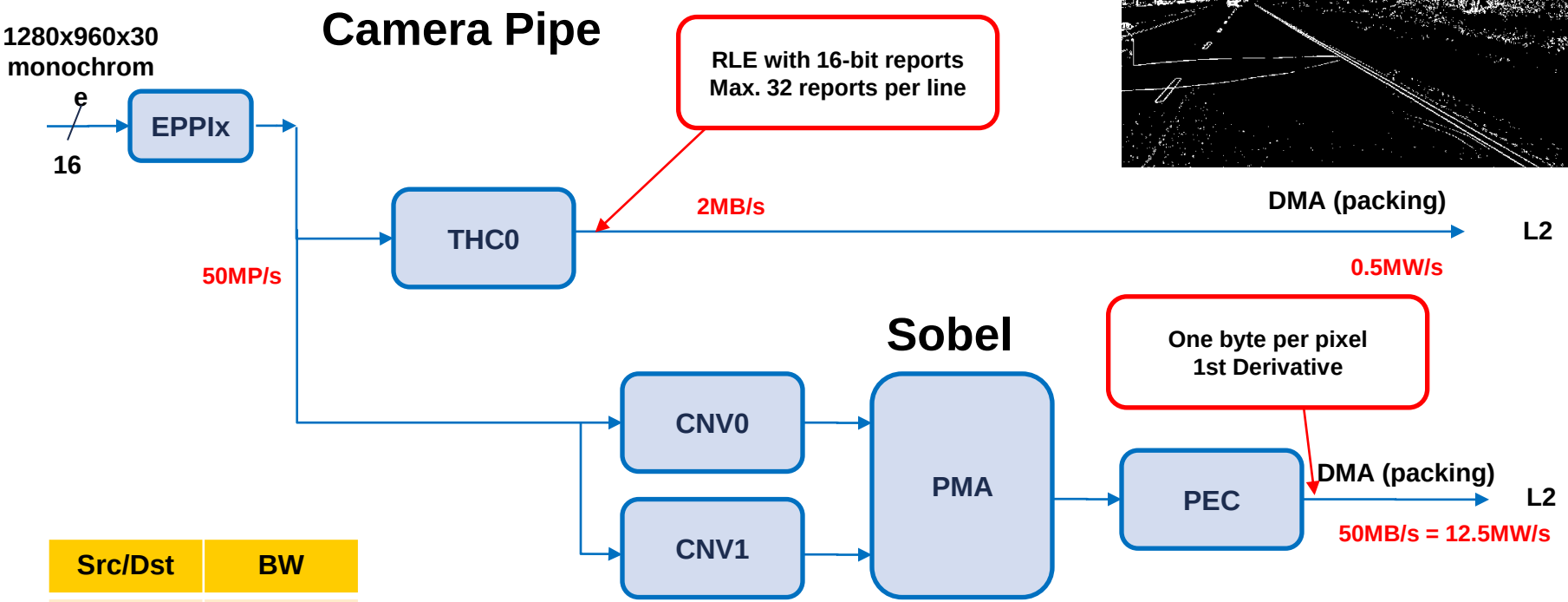
- **Edge Classification/Packing Block**
 - Covers edge enhancement performing non-linear filtering in a pixel neighborhood, edge classification based on orientation, sub-pixel position interpolation
 - Packs the class, vertical/horizontal sub-pixel position into one byte per pixel

- **Threshold/Integral Image Block**
 - 16 x 32-bit threshold function (output => 4bits classification, RLC, rounding up to nearest threshold, finds max. value)
 - Rudimentary histogram function (16 x 32-bit histogram counter, starts relative to the start of frame/line)

ANALOG DEVICES

ADAS Use Case: HD HBLB + LDW (PEC)

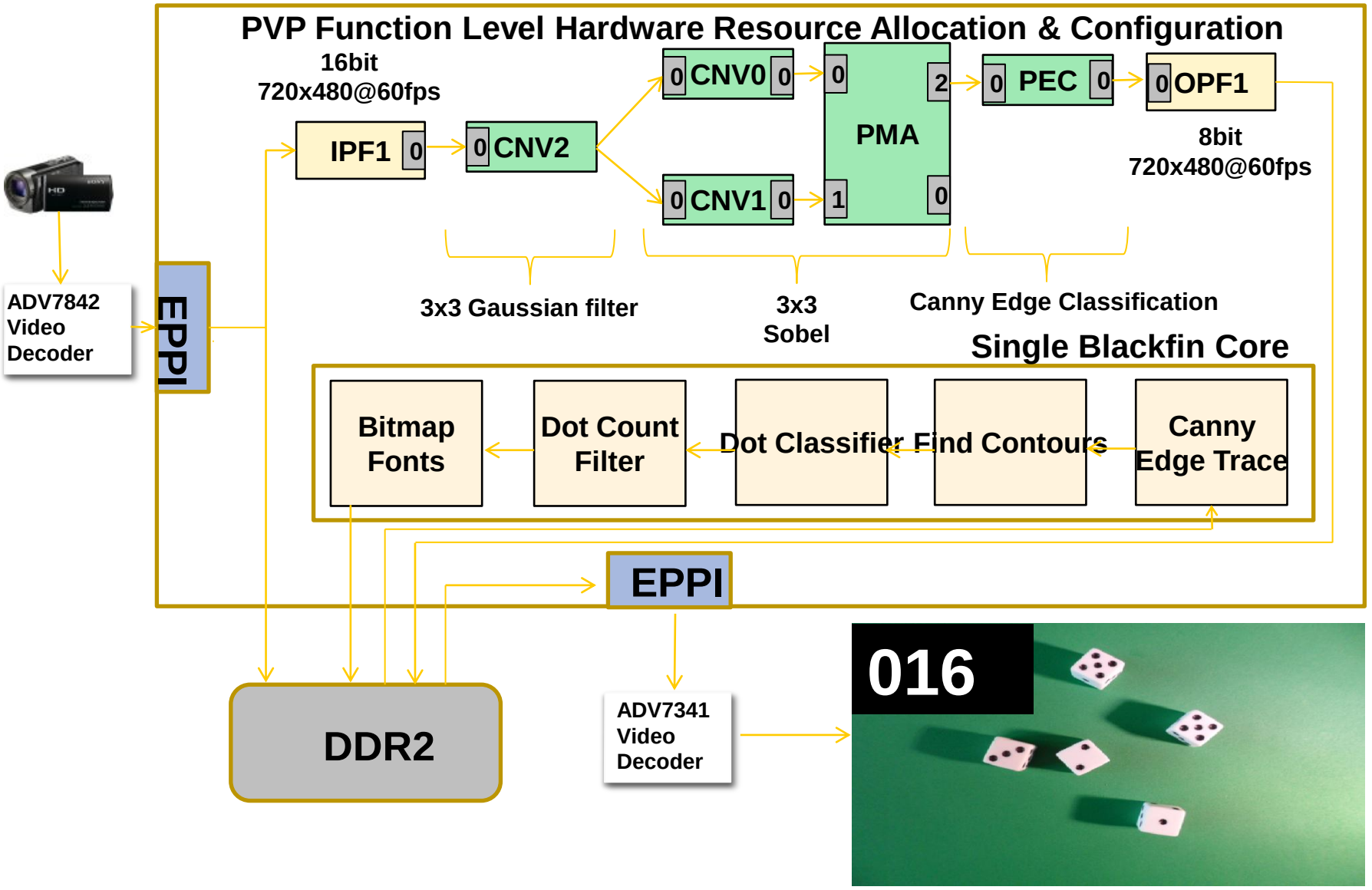




Src/Dst	BW
L3	0 MB/s
L2	52 MB/s
L1	0 MB/s

ANALOG DEVICES

Machine Vision Use Case: Dice Dot Counting



ANALOG DEVICES

High Performance, Parallelism Lower Frequency & Low Power

- **TSMC 65nm GP High Performance Process**
 - 25 MAC ~ 50KGates
 - 5 MAC ~ 25KGates

Convolution Architecture		Power Dissipation		
Number of MACs	Clock Speed (MHz)	Leakage (mW)	Dynamic (mW)	Total (mW)
25 MACs	50	11.3	9.5	**20.7**
5 MACs	250	9.4	17.3	**26.7**

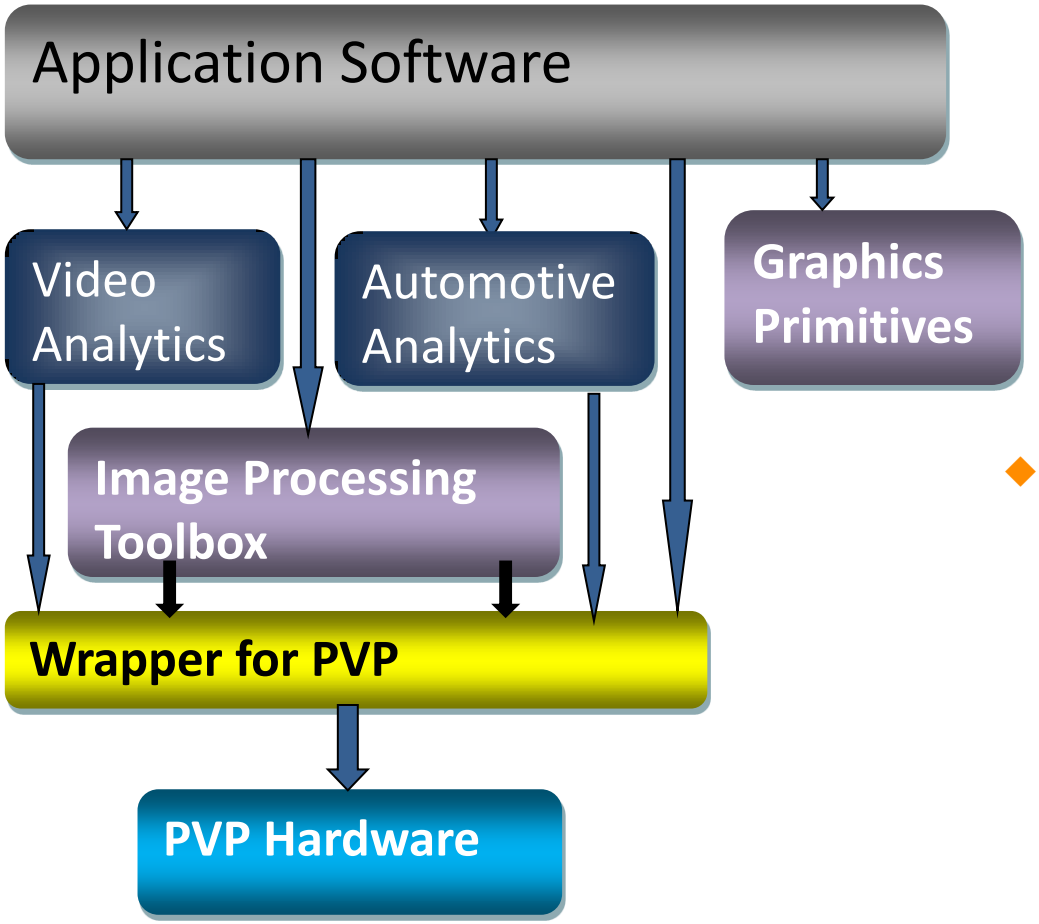
- **TSMC 65nm LP Low Power Process**
 - 25 MAC ~ 55KGates

Convolution Architecture		Power Dissipation		
Number of MACs	Clock Speed (MHz)	Leakage (mW)	Dynamic (mW)	Total (mW)
25 MACs	50	0.2	13.6	**13.8**

- 5 GMACs @ 55mW and ZERO incremental BW due to extensive pipelining at multiple levels of the architecture and optimized function level processing
- 11 mW per GMAC

ANALOG DEVICES

Video/Image Analysis: Software Architecture



- **Blackfin Image Processing Toolbox is a collection of hundreds of optimized functions for image analysis & manipulation.**
 - **Few examples are Histogram operations, morphological operations, 2D convolutions**
- **Video Analytics Toolbox is a set of high level functions that are focused on solving Intelligent Video Surveillance applications.**
 - **Current release supports foreground Object/Blob detection**
 - **Uses Image Processing Toolbox functions**

ANALOG DEVICES

Representative Image Processing, Automotive & Industrial Analytics Toolbox Functionality Available Today (Includes Hardware Mapping as Appropriate)

- Color Conversion
- Image Statistical Tools
- ADAS Modules
- Object & Feature Recognition
- Image Filtering
- Shape-structure Analysis &Computational Geometry
- Geometric Transformations
- Camera Calibration

ANALOG DEVICES

Tradeoffs, Take-aways, & Conclusions (1)

- **An appropriate architectural solution can best be derived from a detailed understanding of market and technical requirements acquired through close customer collaboration and extensive end product technical domain knowledge**
 - **The pipelined vision processor was architected and defined through customer collaboration coupled with general vision and imaging technical hardware and software domain expertise**
- **Hardware/Software/IP partitioning is very important and ultimately determines solution power, performance, and cost**
 - **Choosing to perform appropriate required functions in software on one or more symmetrical or asymmetrical instruction level processors provides many advantages including flexibility**
 - **Partitioning highly computationally complex imaging or vision processing into the appropriate hardware functional IP blocks will generally lead to a cost optimized, low power, and reduced memory bandwidth solution**
 - **Optimizing pixel datapaths and flow in an imaging or vision focused SOC is very important when defining a low cost and power solution**

ANALOG DEVICES

Tradeoffs, Take-aways, & Conclusions (2)

- Many systems on chip architectures will continue to require functionality and IP driven by multiple markets and many applications
 - The BF60x SOC was architected and defined to meet the requirements across multiple markets(e.g. automotive ADAS and industrial vision) and across many applications(e.g. lane departure warning, traffic sign recognition, and barcode reading)
- Trade offs involving instruction level processors, function level processors, dedicated internally developed IP, and 3rd party IP must be weighed carefully in order to arrive at the most optimal SOC architecture and general definition
 - The BF609 contains instruction level digital signal processors, a function level processor which is comprised partly of dedicated internally developed vision focused IP, and 3rd party IP
 - The selection, partitioning, and definition of these SOC components is vital to meeting challenging customer and industry competitive requirements
- Architecting and defining a highly efficient crossbar interconnect and DDR memory controller IP is critical to ensure that the system meets all of the bandwidth and latency requirements across many demanding masters
 - Arbitration and prioritization optimization throughout the entire data path from master to slave is paramount to satisfying all master requirements when executing highly computationally complex vision applications

ANALOG DEVICES



The World Leader in High Performance Signal Processing Solutions

- Email
 - Robert.Bushey@analog.com
- www.analog.com/BlackfinModules
 - Vision Analytics Toolbox(VAT)
 - Image Processing Toolbox(IPTBX)
 - ADAS Vision Analytics Toolbox(AVAT)
 - 2D Graphics Libraries(BF2DGL)
- www.analog.com/Blackfin
 - Blackfin Processors & SOCs
- automotive.analog.com
 - Automotive and ADAS

ANALOG DEVICES

TOSHIBA

Leading Innovation

Visconti2 - A Heterogeneous Multi-Core SoC for Image-Recognition Applications

Masato Uchiyama, Hideho Arakida, Yasuki Tanabe,
Tsukasa Ike, Takanori Tamai, Moriyasu Banno

Toshiba Corporation, Kawasaki, Japan

Copyright 2012, Toshiba Corporation.

Outline

- **Background**
- **Visconti2**
 - Overview of architecture and chip
 - CoHOG accelerator

 (Co-occurrence Histograms of Oriented Gradients)
- **Real Applications**
 - Monocular Pedestrian Detection
 - Hand Gesture User Interface (UI)
- **Conclusion**

© 2012 Toshiba Corporation

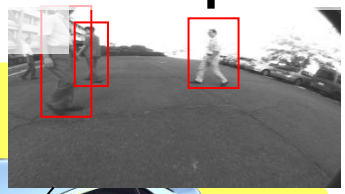
Background: Targets of Visconti2

Image recognition technology ⇨ A variety of products

Forward collision warning



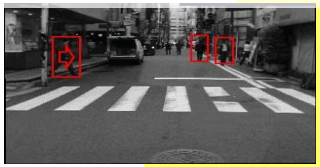
Backover prevention



Door security



Pedestrian detection

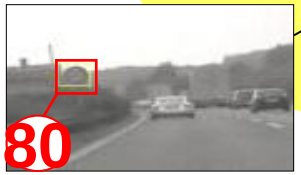




Driver monitoring



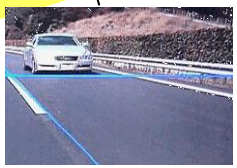
Traffic sign recognition



Face tracking for glassless 3D



Lane change assistance



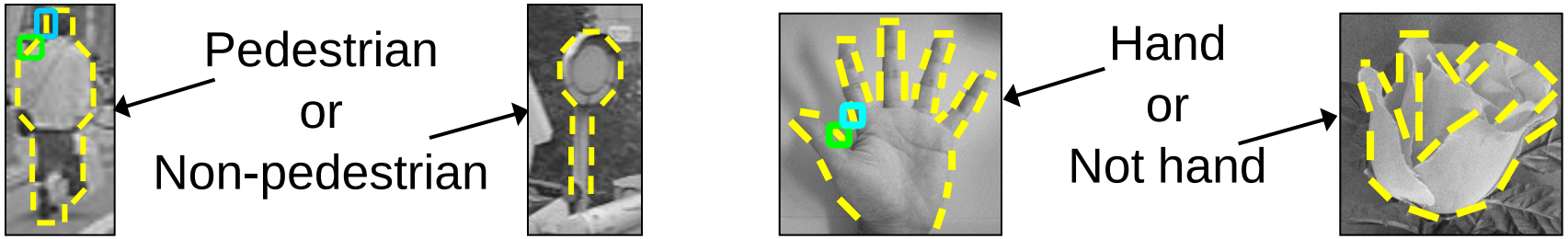
Visconti2 designed for

- **Automotive :** Advanced Driver Assistance Systems (ADAS)
- **Consumer**
- **Industry**

© 2012 Toshiba Corporation

Background: Requirements & Approach

- **High accuracy of object recognition**



CoHOG (Co-occurrence Histograms of Oriented Gradients)
- One of the most accurate image feature descriptors
- Toshiba original (T.Watanabe et al., Proc. PSIVT 2008, pp.37-47)

- **High performance**
 - E.g. Monocular Pedestrian Detection **using CoHOG**
 ➔ 3,983ms/frame on 1GHz CPU

40x speedup required by real-time execution

- **Low power consumption**
 - Cooling without fan (< 1W in typical condition)

Hardware accelerators for frequently used tasks which are performance bottlenecks (CoHOG, etc.)

© 2012 Toshiba Corporation

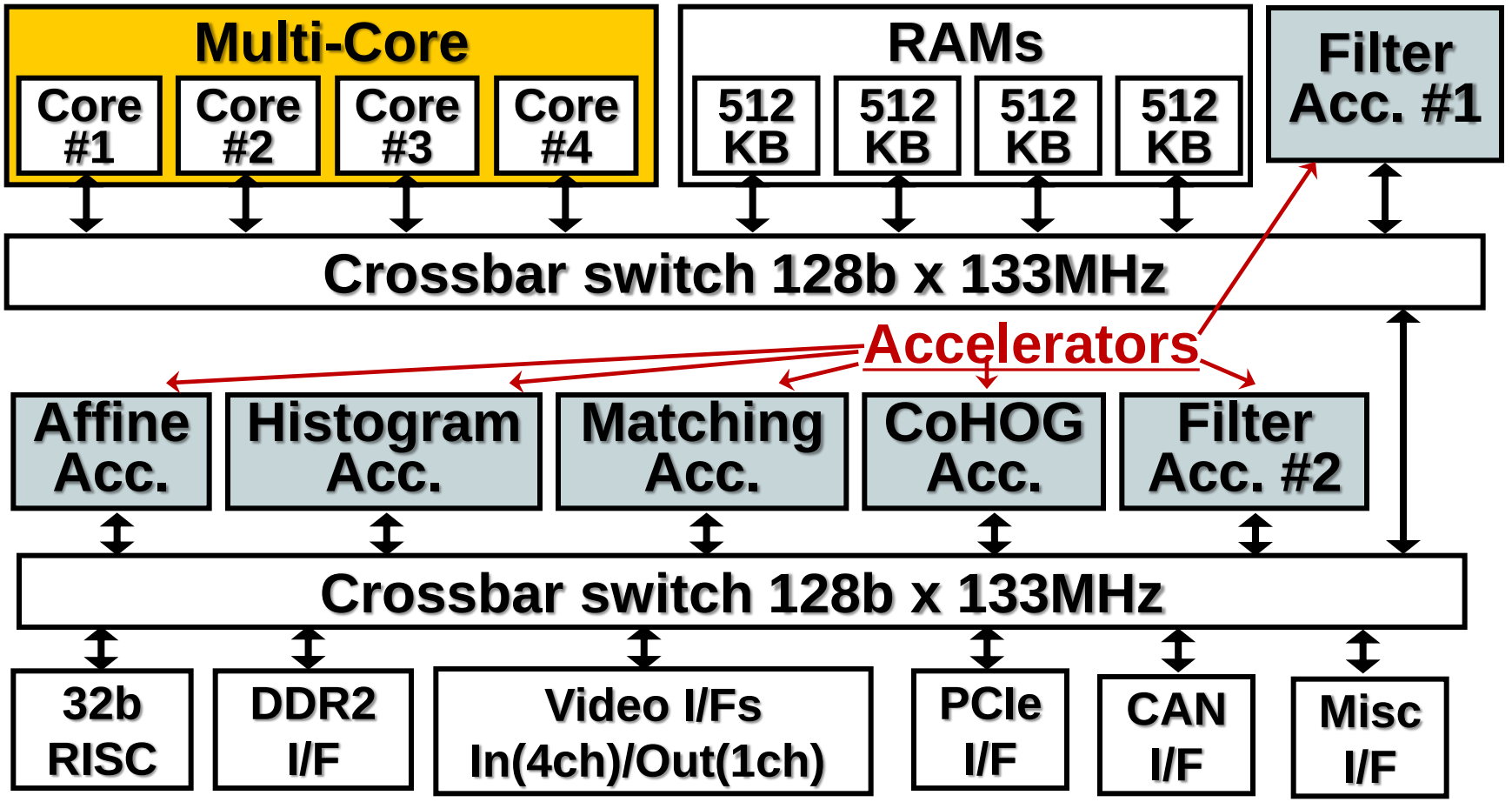
Outline

- **Background**
- **Visconti2**
 - Overview of architecture and chip
 - CoHOG accelerator

 (Co-occurrence Histograms of Oriented Gradients)
- **Real Applications**
 - Monocular Pedestrian Detection
 - Hand Gesture User Interface (UI)
- **Conclusion**

© 2012 Toshiba Corporation

Chip Architecture



Memory Bandwidth

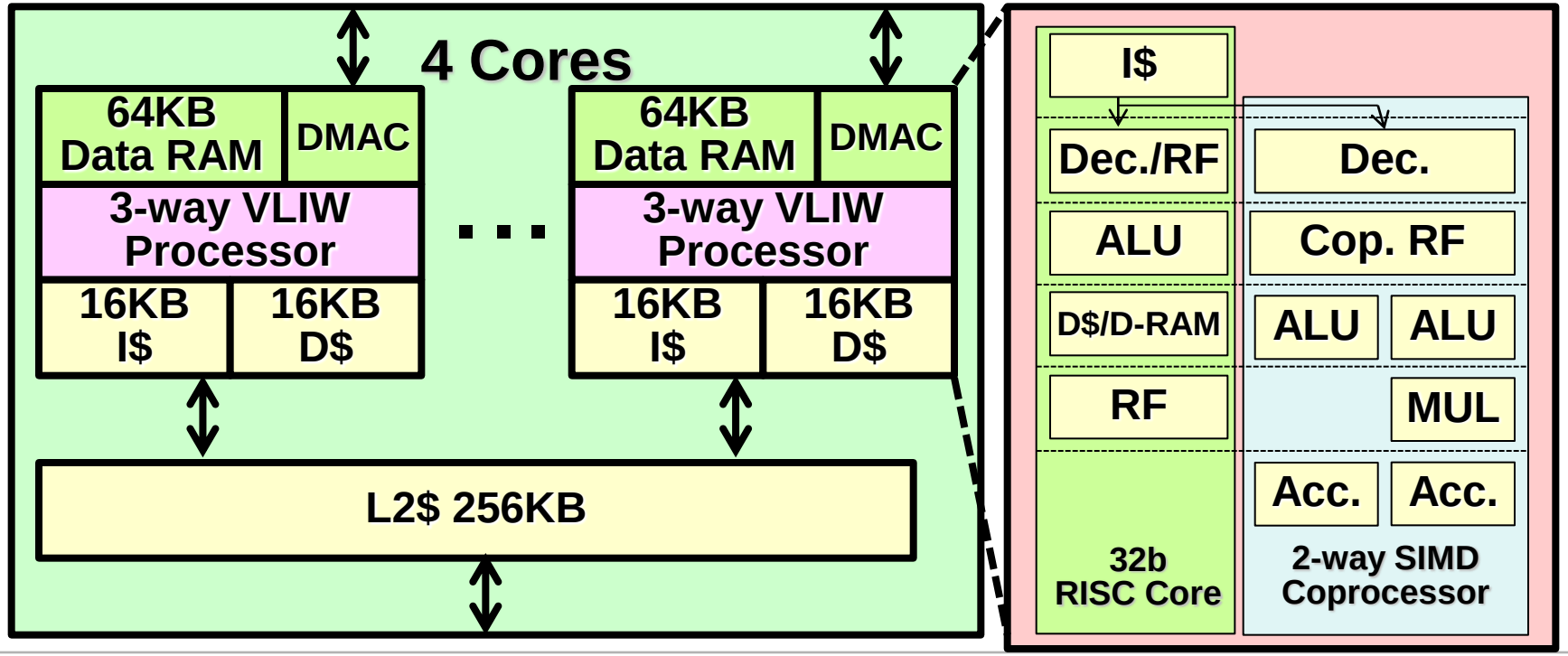
DDR2: Peak 2GB/sec

On-chip RAMs: 2GB/sec x 4ch.

© 2012 Toshiba Corporation

Multi-core Subsystem

- **Four homogeneous VLIW cores with 256KB L2$**
 - **3-way VLIW core**
 - **RISC core + 2-way SIMD coprocessor (ISSCC '08[S.Nomura])**
 - **Additional 64KB data RAM and DMA controller**
 - **Exploit multi-grain parallelism**
 - **Application, task and thread level parallelism: by four cores**
 - **Data level parallelism: by SIMD coprocessor**

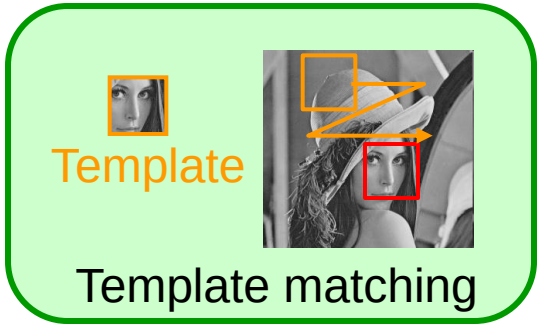


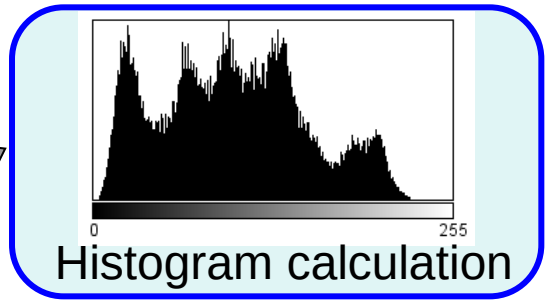
© 2012 Toshiba Corporation

Hardware Accelerators

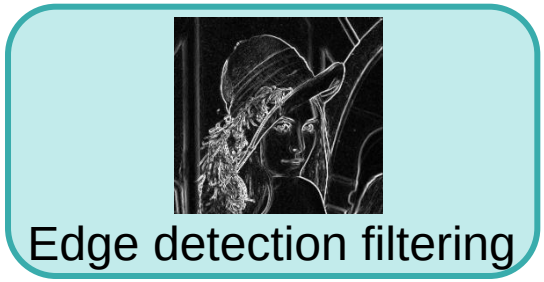
- Six accelerators implemented
 - **CoHOG accelerator**
 - **Matching accelerator**
 - **Histogram accelerator**
 - **Affine accelerator**
 - **Two Filter accelerators**











Realizing
"High performance with
low power consumption"

➔ **We adopted "Highly parallelized" approach rather than "High clock frequency" approach.**

© 2012 Toshiba Corporation

CoHOG based Recognition

- **Extension to widely-used HOG (Histogram of Oriented Gradients)**

1. Make gradient orientation image

Region of Interest (ROI)

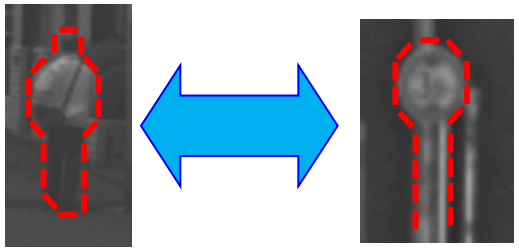






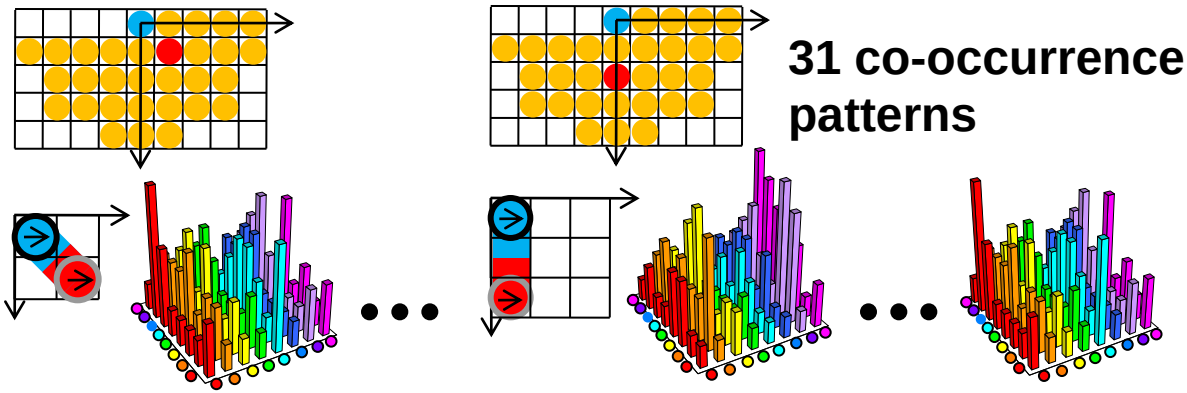


Similar on HOG

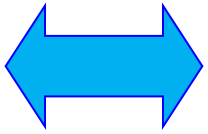


Different on CoHOG

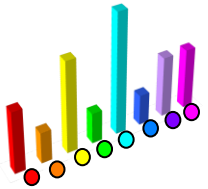
2. Calculate co-occurrence histogram



31 co-occurrence patterns



HOG



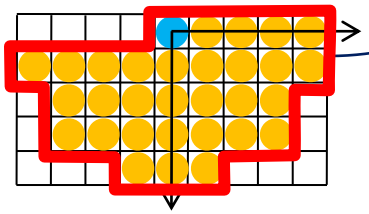
8 gradient orientations

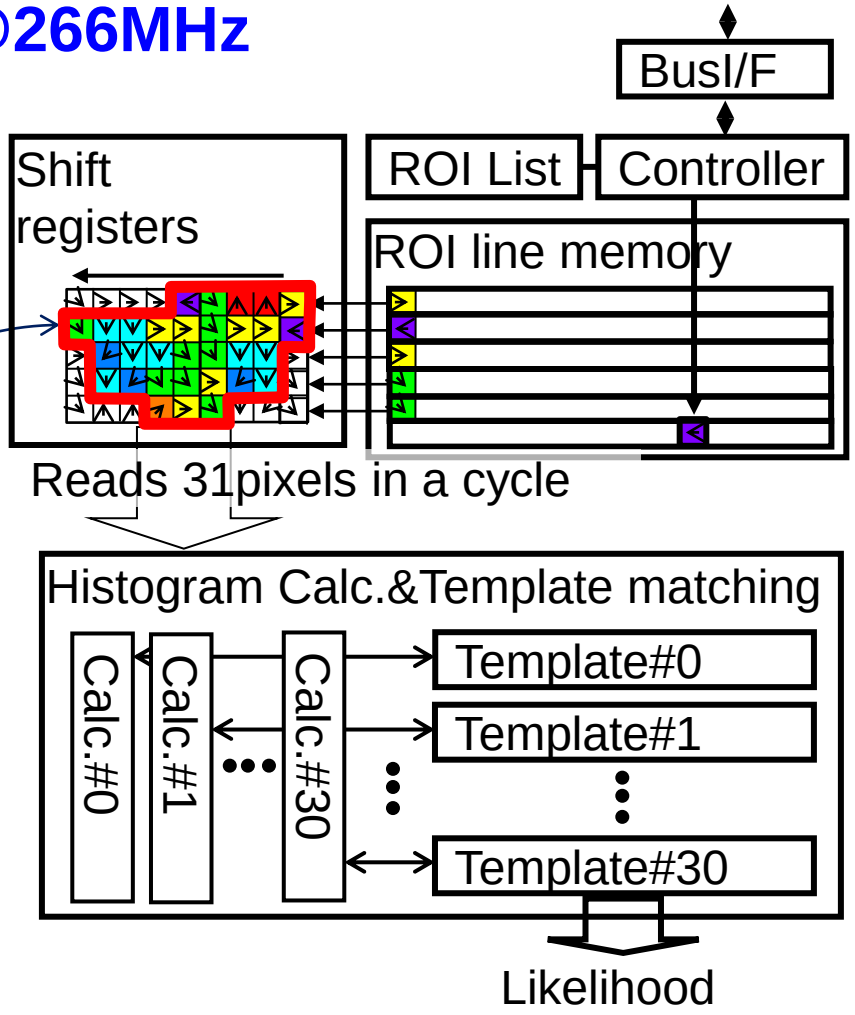
Higher accuracy

© 2012 Toshiba Corporation

CoHOG Accelerator

- **Throughput: 1 pixel / clock @266MHz**
 - **31 co-occurrence pairs are calculated in a clock cycle.**
 - **31 x 3 arithmetic operations**
 - **31 x 2 data references**
 - **Pixel range check**





Over 400,000 ROIs/sec

(18 x 36 pixels/ROI)



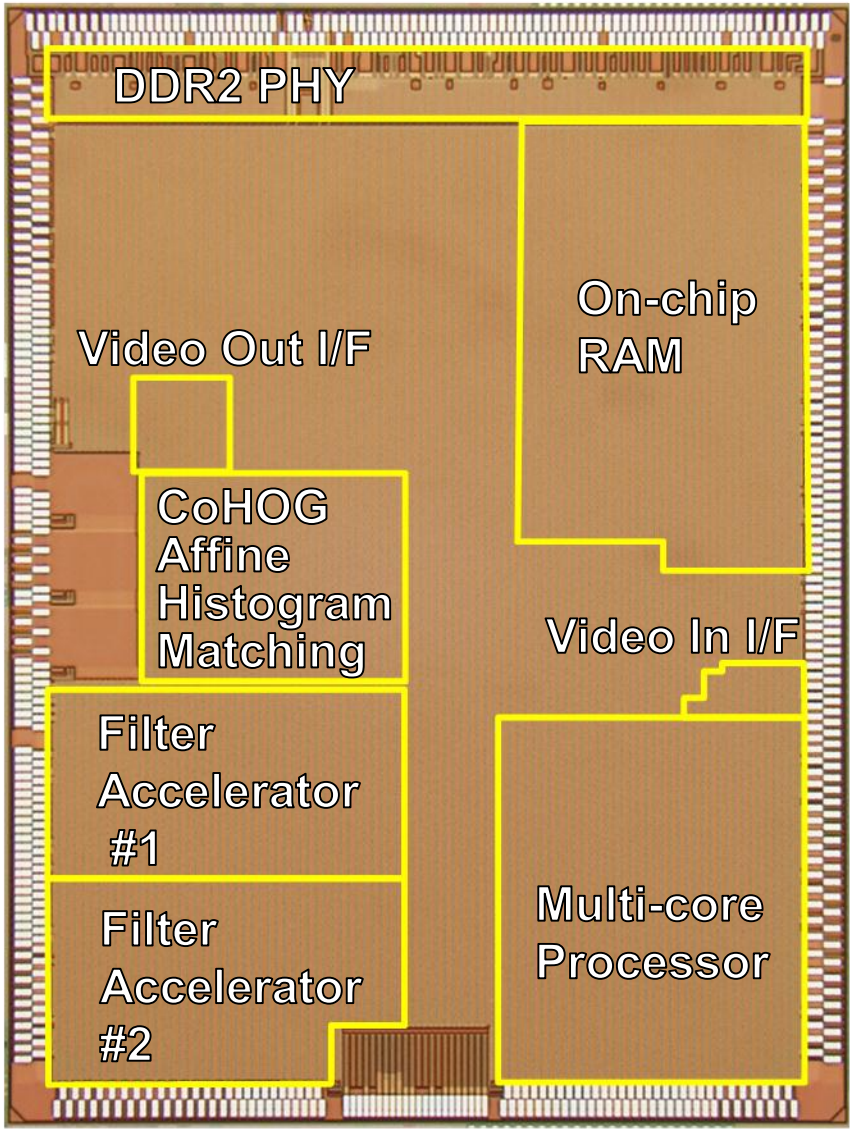
400,000 ROIs/sec is enough for our target applications.

© 2012 Toshiba Corporation

Features and Chip Micrograph

Process	**40nm**
Chip Size	**44.54mm^2**
Supply Voltages	
Core	**1.1V**
DDR2/PCIe PHY	**1.8V**
I/O	**3.3V**
Performance	
Total peak performance	**464GOPS**
Power efficiency	**620GOPS/W**

(Y.Tanabe et al., Proc. ISSCC 2012, pp.222-223)



© 2012 Toshiba Corporation

Outline

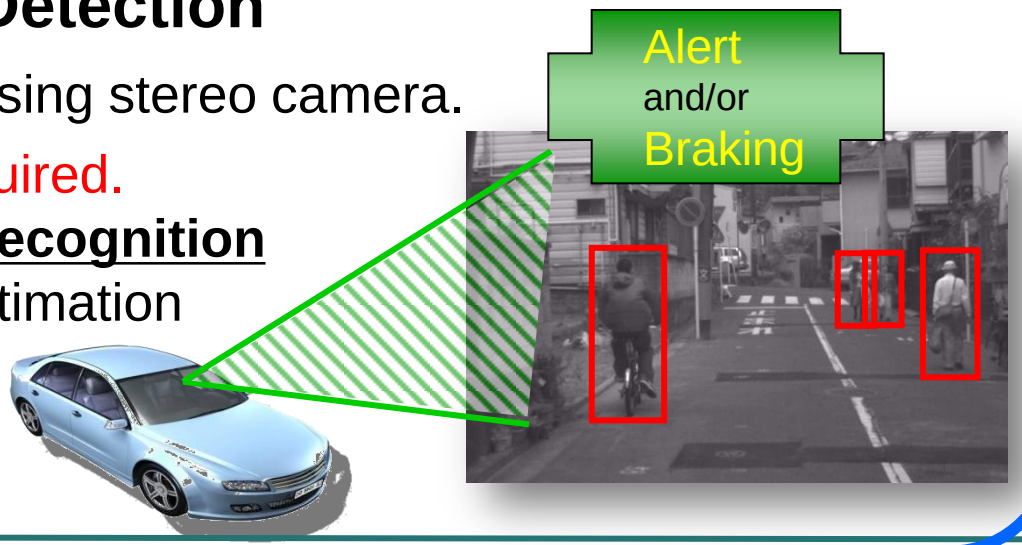
- **Background**
- **Visconti2**
 - Overview of architecture and chip
 - CoHOG accelerator

 (Co-occurrence Histograms of Oriented Gradients)
- **Real Applications**
 - Monocular Pedestrian Detection
 - Hand Gesture User Interface (UI)
- **Conclusion**

© 2012 Toshiba Corporation

Real Applications

- **Monocular Pedestrian Detection**
 - System cost is lower than using stereo camera.
 - Huge computations are required.
 (**Sliding window CoHOG recognition** is used instead of depth estimation based on stereo matching with stereo camera.)



- **Hand Gesture UI**
 - Hand recognition is applied to many ROIs (**sliding window CoHOG recognition**).
 - High frame rate is required.

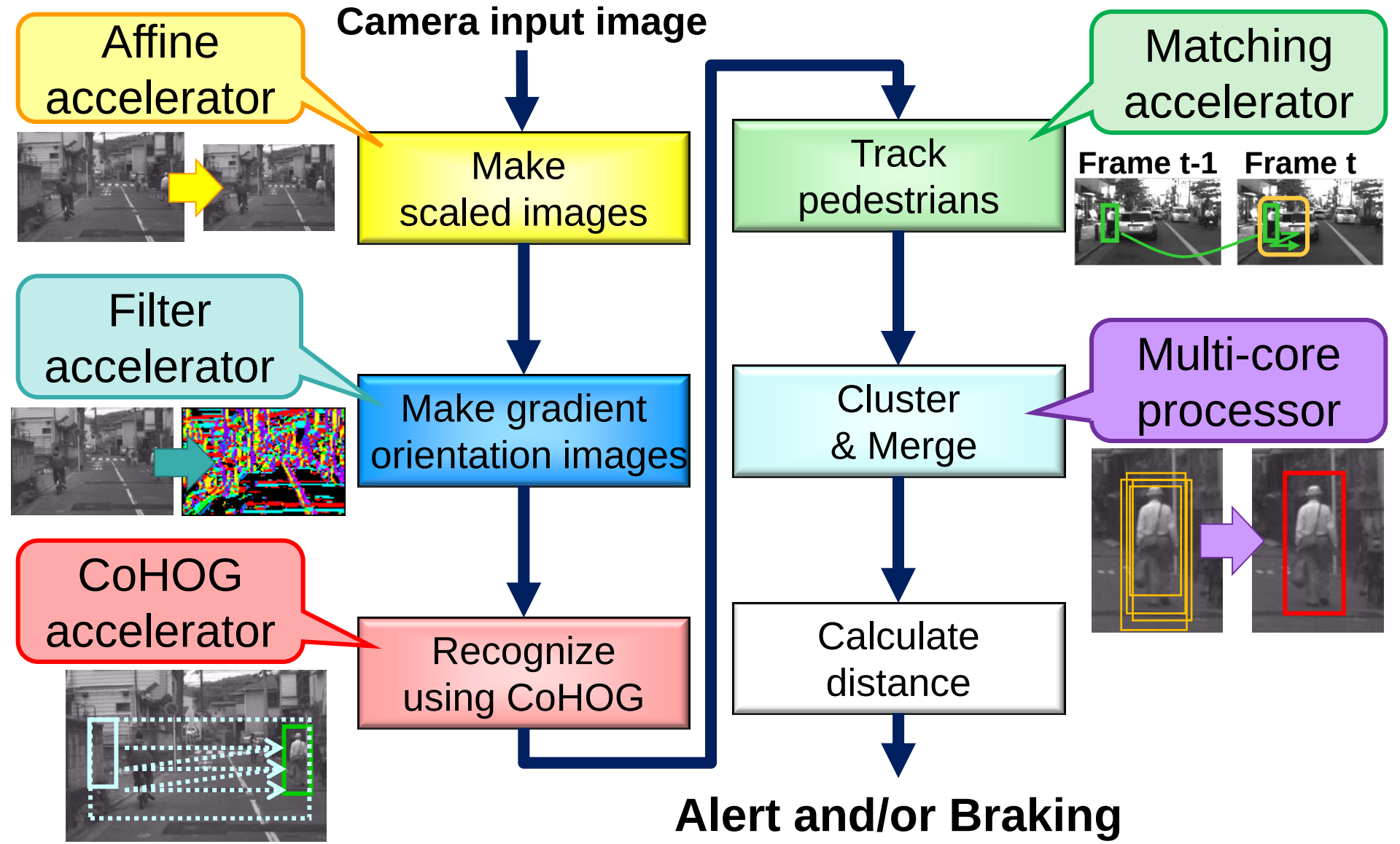
Command examples





© 2012 Toshiba Corporation

Pedestrian Detection : Processing Flow



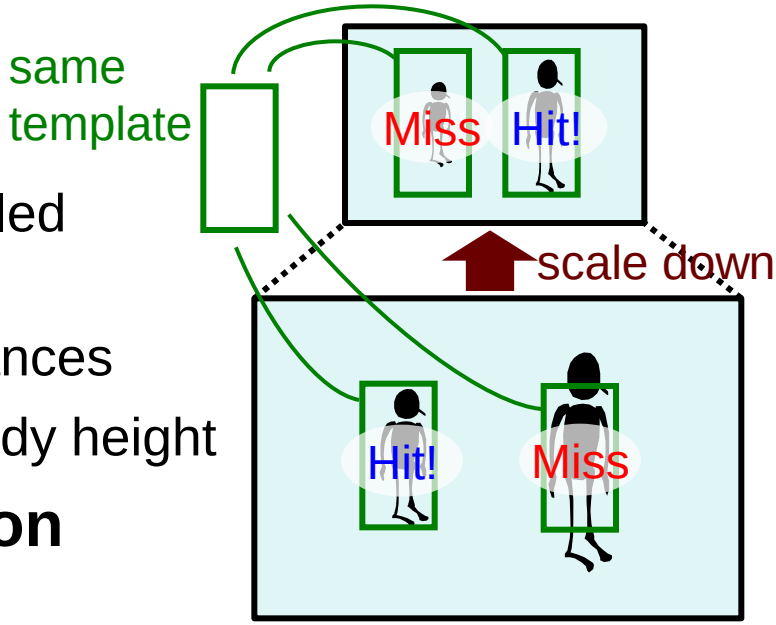
© 2012 Toshiba Corporation

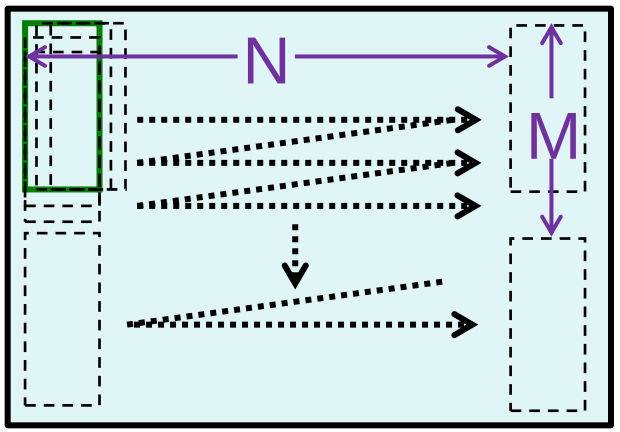
Pedestrian Detection : CoHOG Recognition

- **A number of scaled images are generated by Affine accelerator.**
 - A template is used to match with the scaled images:
 - To detect pedestrians in different distances
 - To detect pedestrians with different body height
- **Sliding window CoHOG recognition**

 ➔ 650 ROIs / image @ VGA
- **Performance requirement of CoHOG recognition**

 500 (sliding window ROIs on average)
 x 20 (scaled images)
 x 10 (frame / sec)
 = 100,000 ROIs/sec
 < CoHOG accelerator : 400,000 ROIs/sec

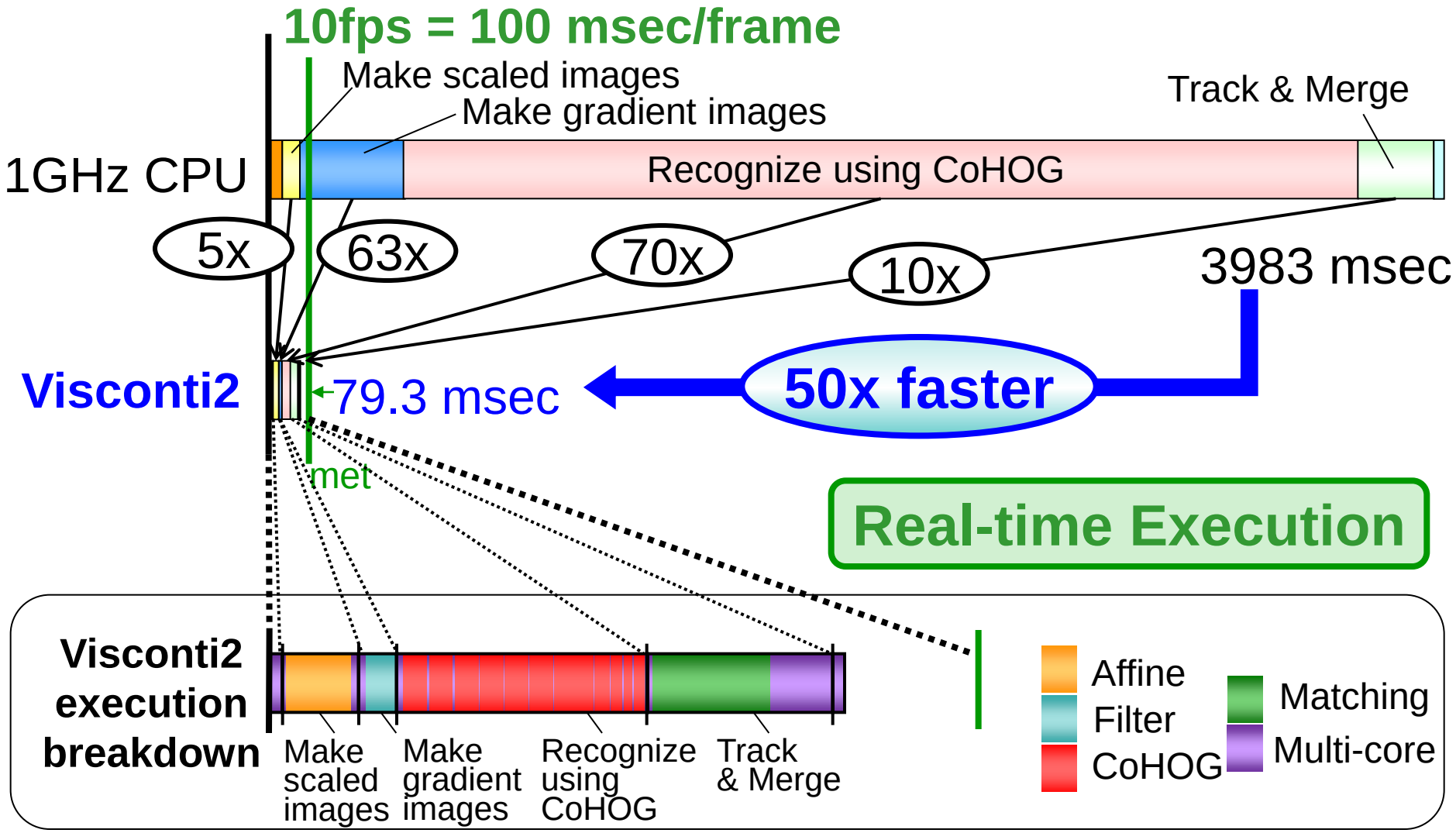




© 2012 Toshiba Corporation

Pedestrian Detection : Execution Time

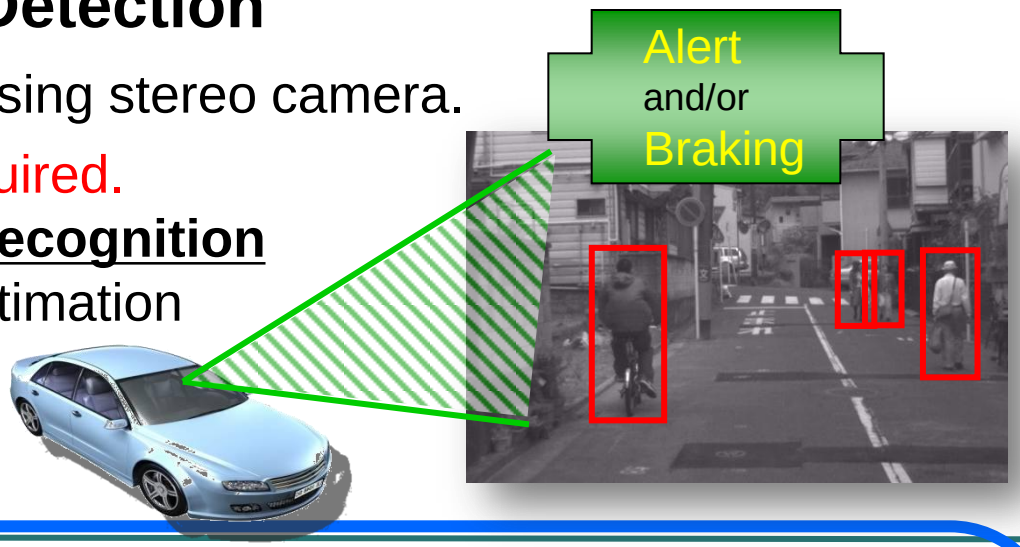
- **Execution time per frame**



© 2012 Toshiba Corporation

Real Applications

- **Monocular Pedestrian Detection**
 - System cost is lower than using stereo camera.
 - Huge computations are required.
 (**Sliding window CoHOG recognition** is used instead of depth estimation based on stereo matching with stereo camera.)



- **Hand Gesture UI**
 - Hand recognition is applied to many ROIs (**sliding window CoHOG recognition**).
 - High frame rate is required.

Command examples

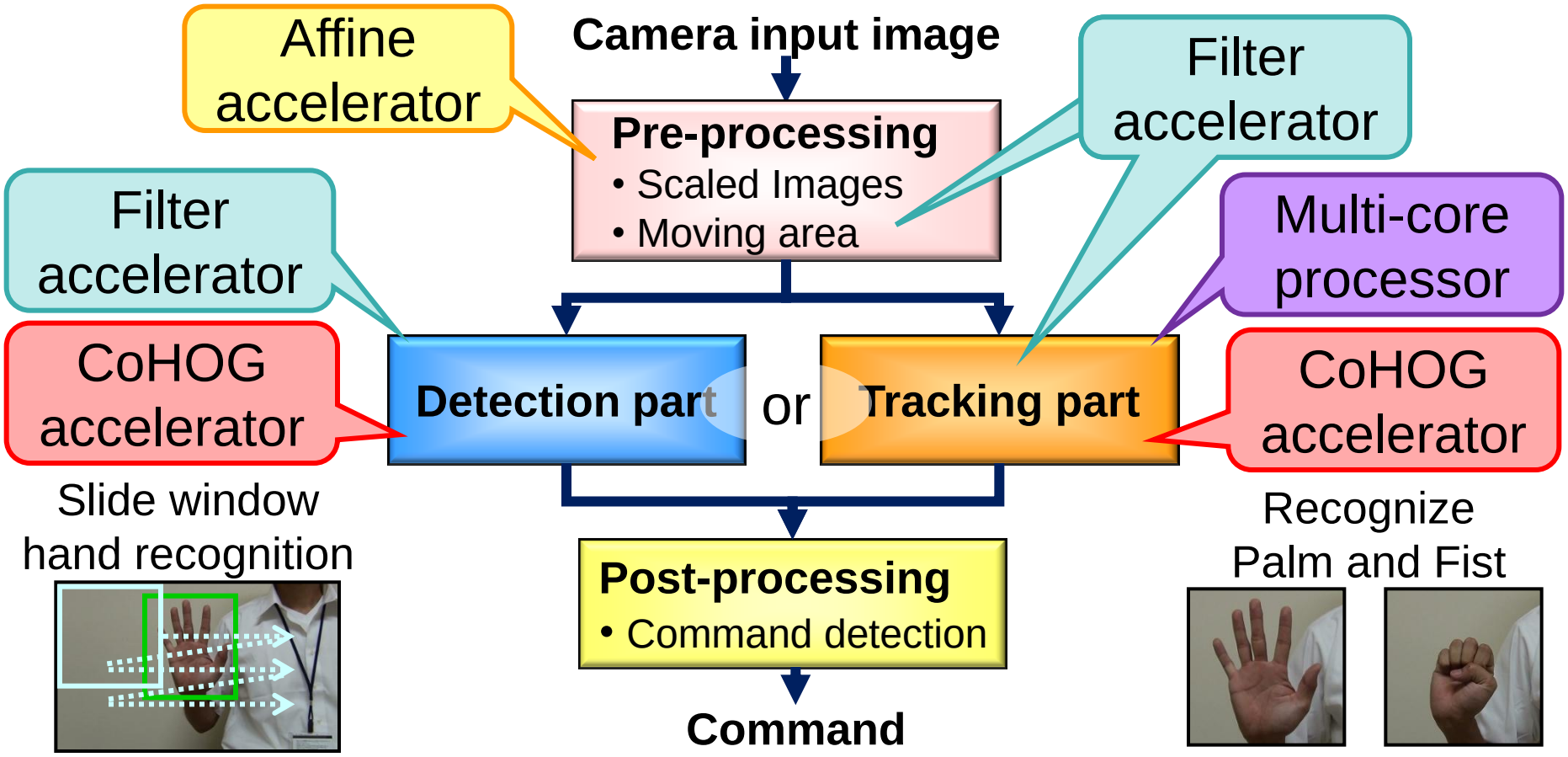




© 2012 Toshiba Corporation

Hand Gesture UI : Processing Flow

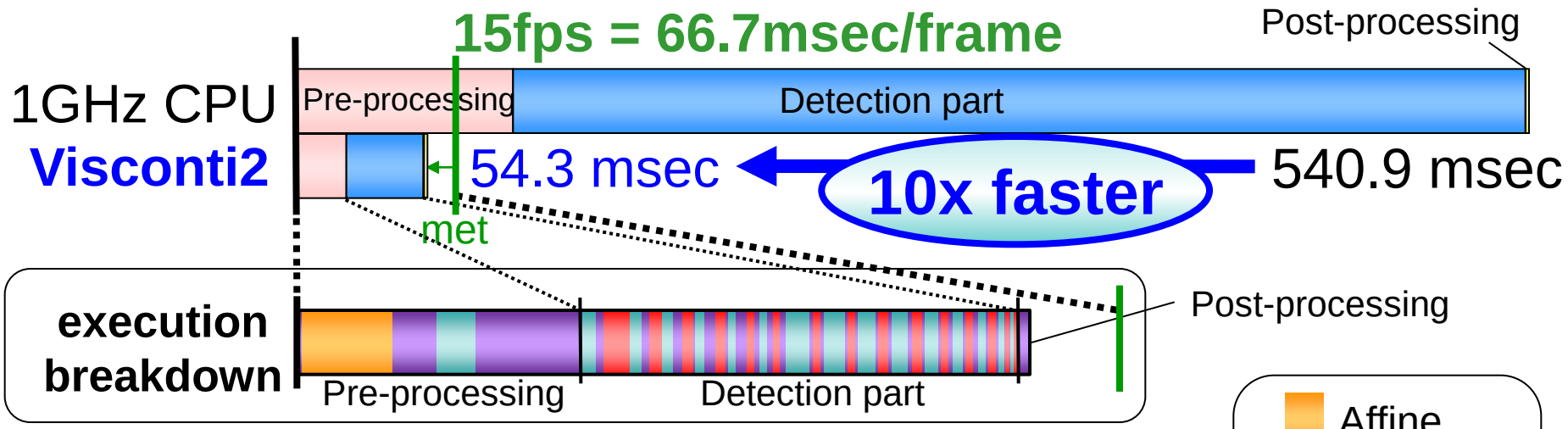
- **Switching between two processing modes**
 - **Detection mode** : sliding window hand recognition @ 15fps
 - **Tracking mode** : trajectory recognition @ 30fps



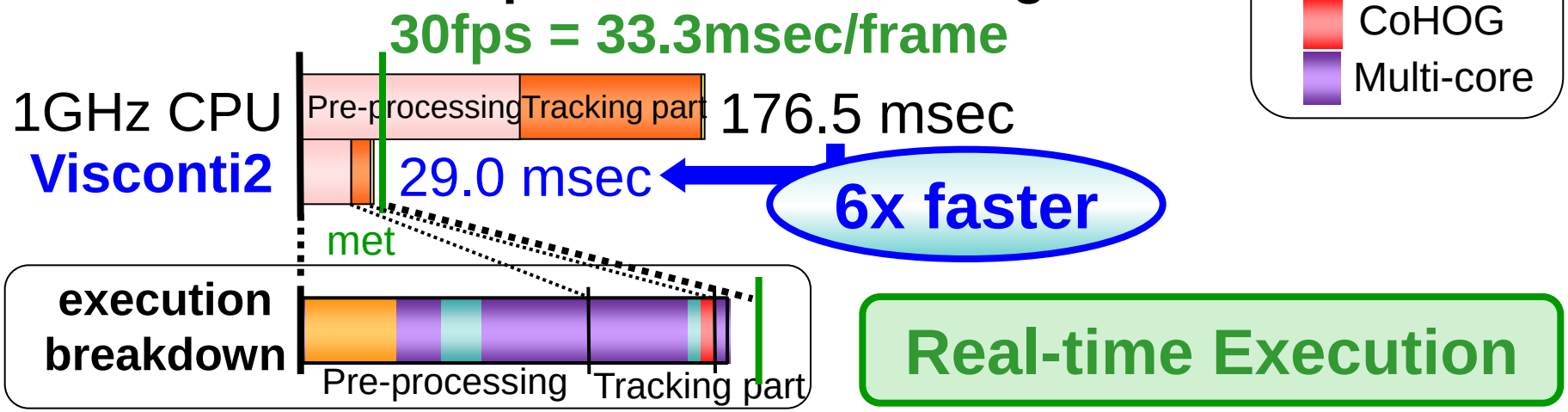
© 2012 Toshiba Corporation

Hand Gesture UI : Execution Time

- **Execution time per frame in detection mode**



- **Execution time per frame in tracking mode**



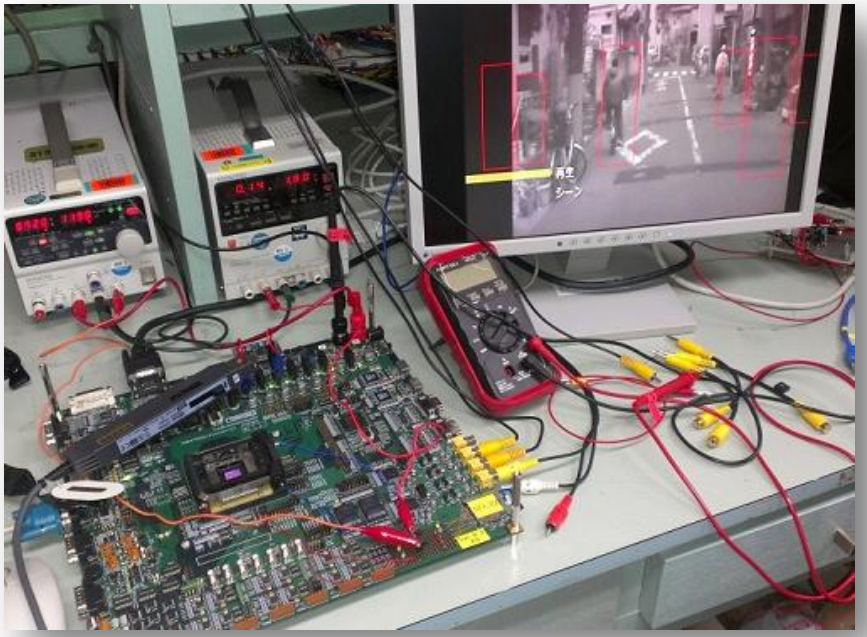
© 2012 Toshiba Corporation

Evaluation of Power Consumption

- **Monocular Pedestrian Detection**
 - Chip total : 870mW
 - Core (1.1V) : 356mW
 - PHY(1.8V) : 460mW
 - I/O (3.3V) : 54mW
- **Hand Gesture UI**
 - Chip total : 891mW
 - Core (1.1V) : 363mW
 - PHY(1.8V) : 472mW
 - I/O (3.3V) : 56mW

< 1W : Cooling without fan

Typical condition:
Process center sample, 25ºC



Evaluation board and
power measurement environment

© 2012 Toshiba Corporation

Conclusion

- **Visconti2 is a heterogeneous multi-core SoC dedicated for image recognition.**

 Visconti2 achieves:

 – Accurate recognition
 - CoHOG based image recognition is implemented.

 – High performance with low power consumption
 - We implemented six highly parallelized hardware accelerators.
 - Under 1W power consumption is achieved. (typical condition)

- **Two real applications on Visconti2 using HW accelerators are demonstrated.**

 – Monocular Pedestrian Detection

 – Hand Gesture User Interface

- **Visconti2 status: ES ready**

 http://www.semicon.toshiba.co.jp/eng/product/assp/selection/automotive/infotain/visconti/

© 2012 Toshiba Corporation

TOSHIBA

Leading Innovation

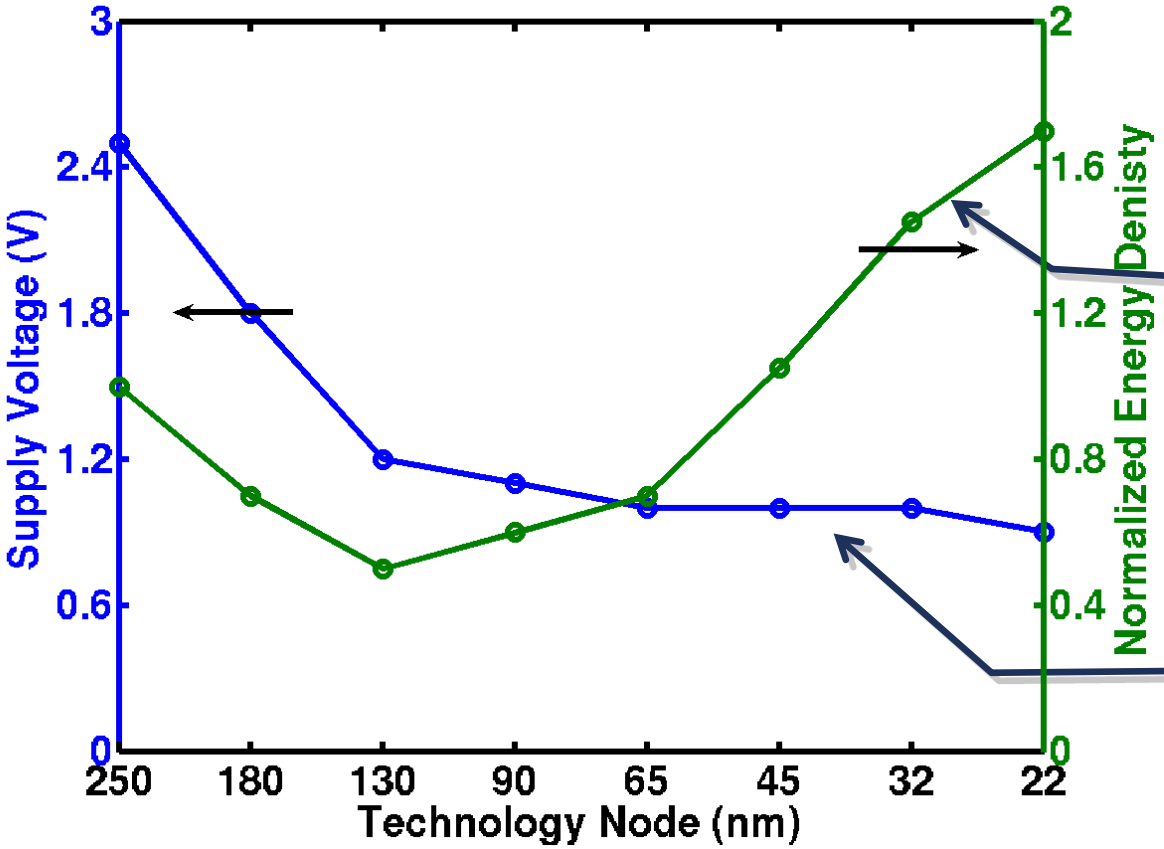
Centip3De: A 64-Core, 3D Stacked, Near-Threshold System

Ronald G. Dreslinski

David Fick, Bharan Giridhar,
Gyouho Kim, Sangwon Seo, Matthew Fojtik,
Sudhir Satpathy, Yoonmyung Lee, Daeyeon Kim,
Nurrachman Liu, Michael Wieckowski, Gregory Chen,
Trevor Mudge, Dennis Sylvester, David Blaauw

University of Michigan

The Problem of Power



Power does not decrease at the same rate that transistor count increases, resulting in increased energy density

Circuit supply voltages are no longer scaling…

Dynamic dominates

$$U \approx \frac{CV_{dd}^2}{A} + \frac{I_{leak}V_{dd}}{Af}$$

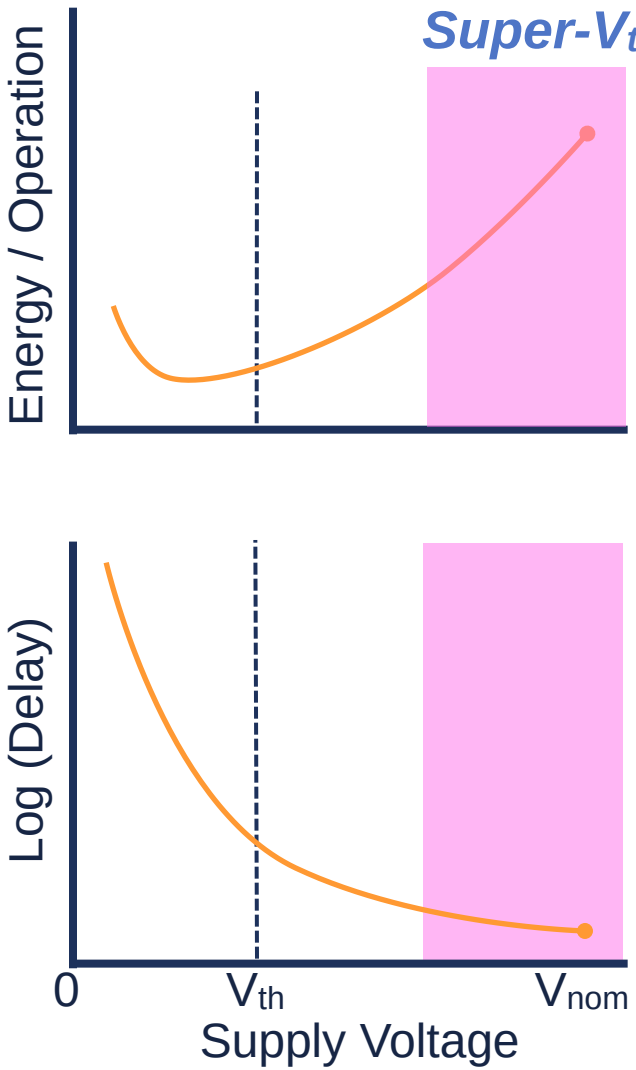
A = gate area → scaling $1/s^2$

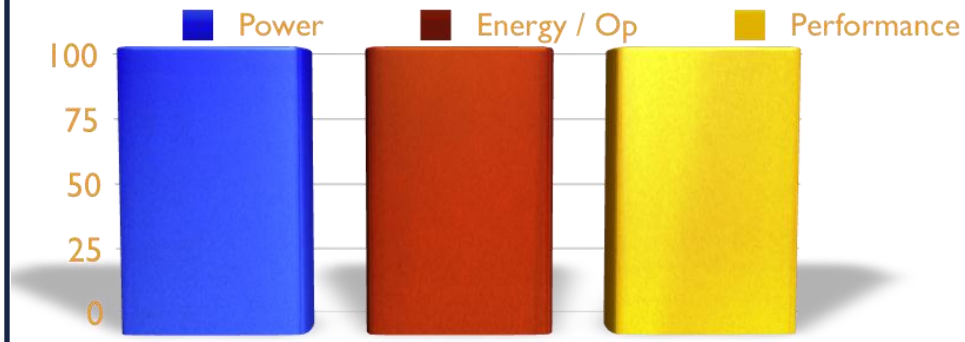
C = capacitance → scaling < 1/s

The emerging dilemma:

More and more gates can fit on a die,
but cooling constraints are restricting their use

Today: Super-V_{th}, High Performance, Power Constrained





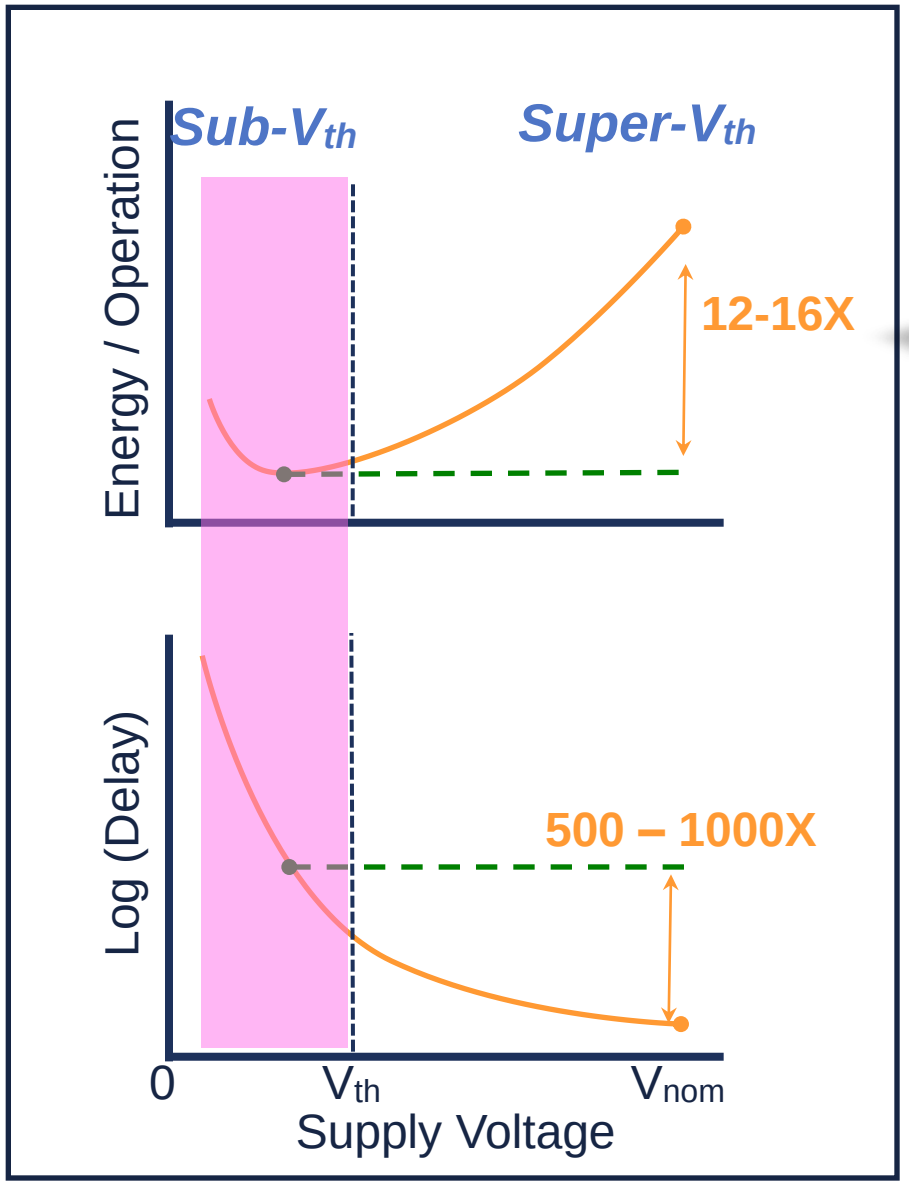
Normalized CPU Metrics

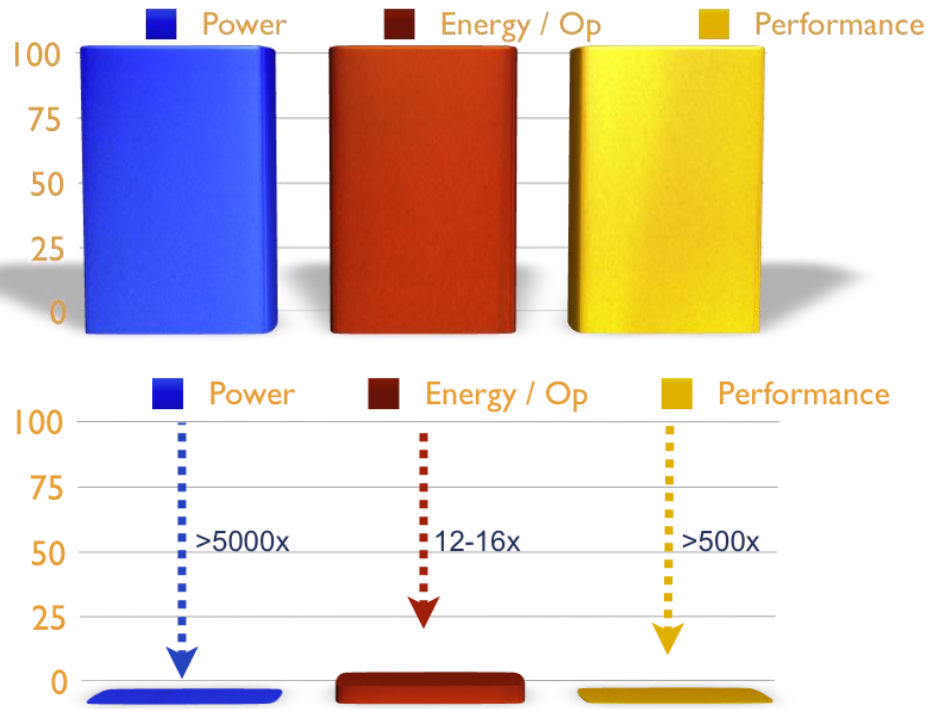
Large gate overdrive favors performance with unsustainable power density

Must design within fixed TDP

Goal: maintain performance, improved Energy/Operation

Subthreshold Design

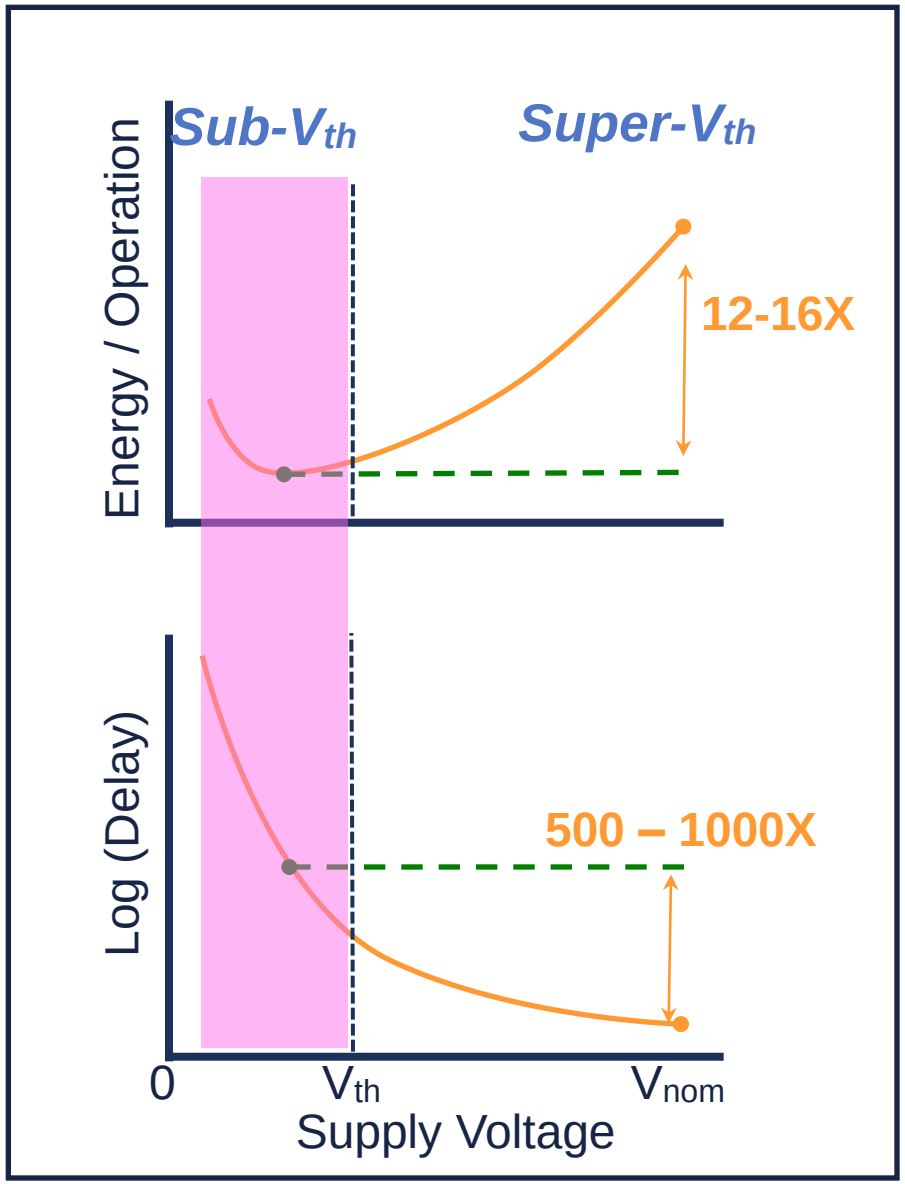


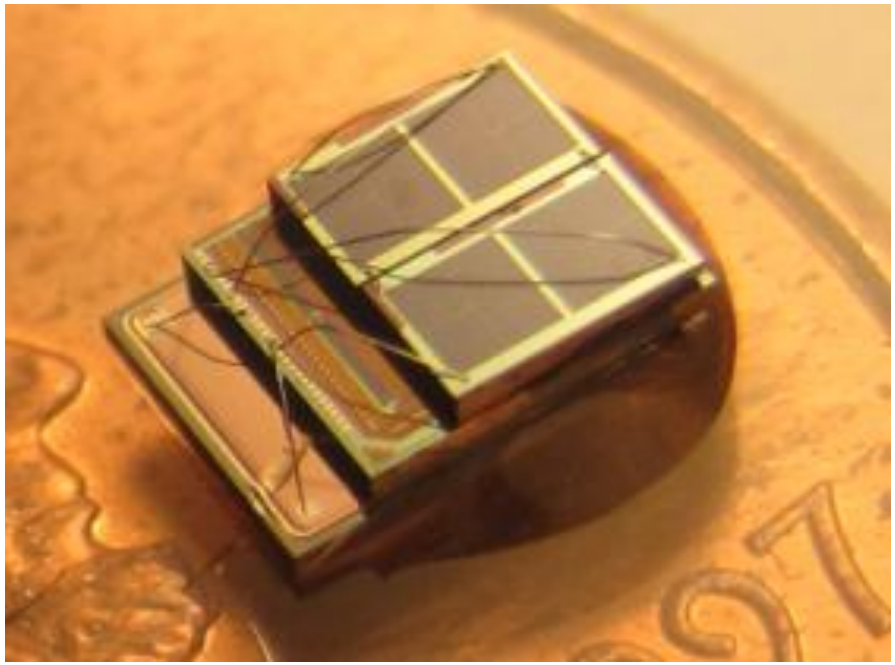


Operating in sub-threshold yields large power gains at the expense of performance.

Applications: sensors, medical

Subthreshold Design



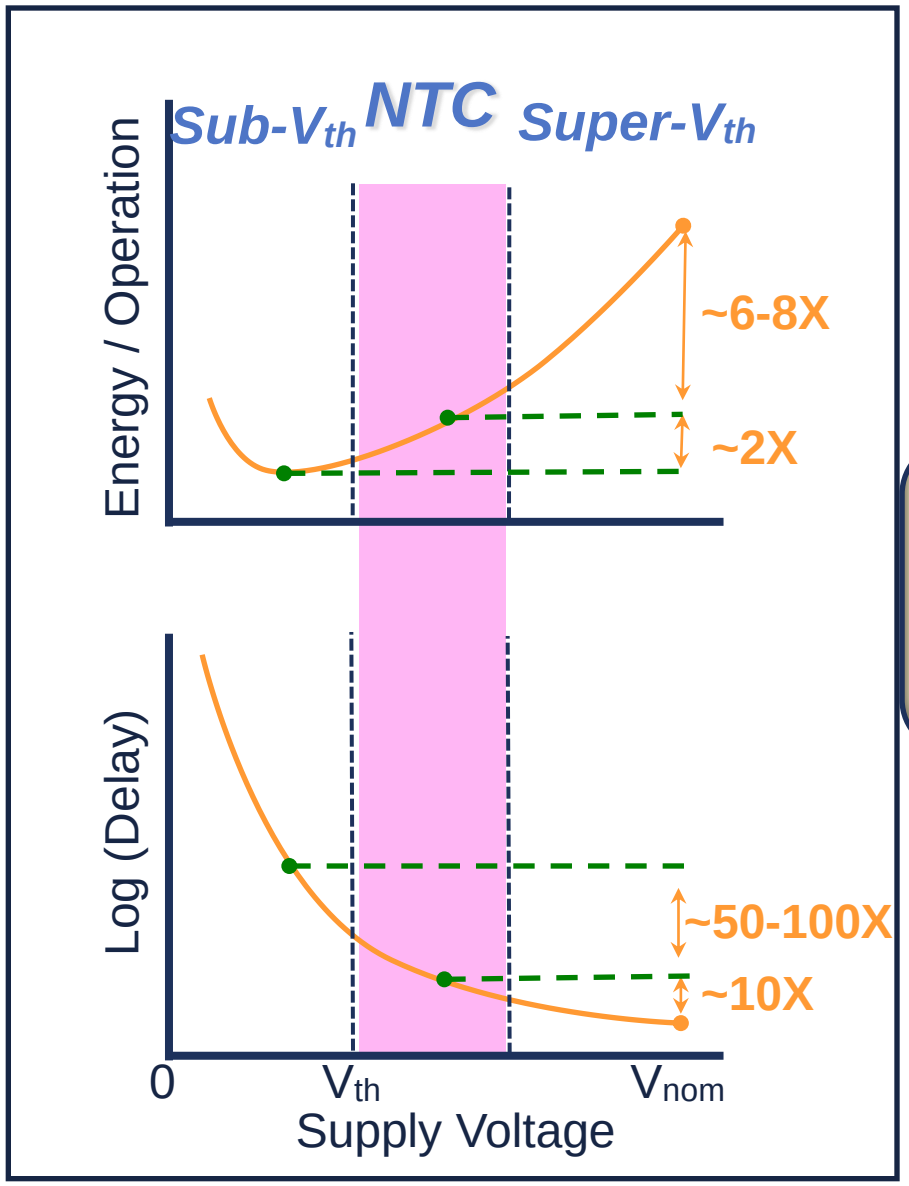


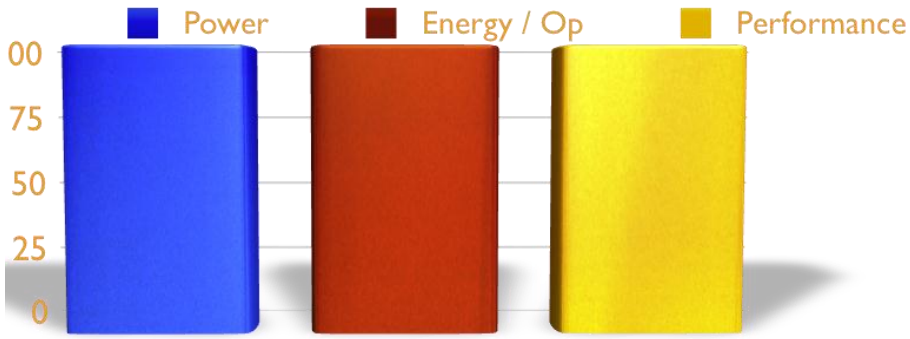
Phoenix 2 Processor, ISSCC'10

Operating in sub-threshold yields large power gains at the expense of performance.

Applications: sensors, medical

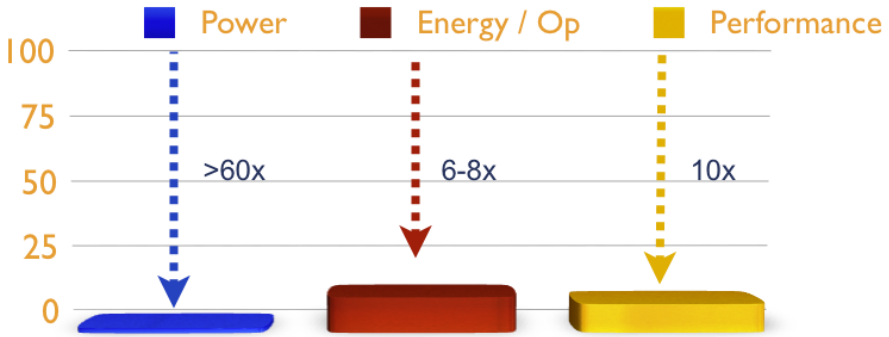
Near-Threshold Computing (NTC)



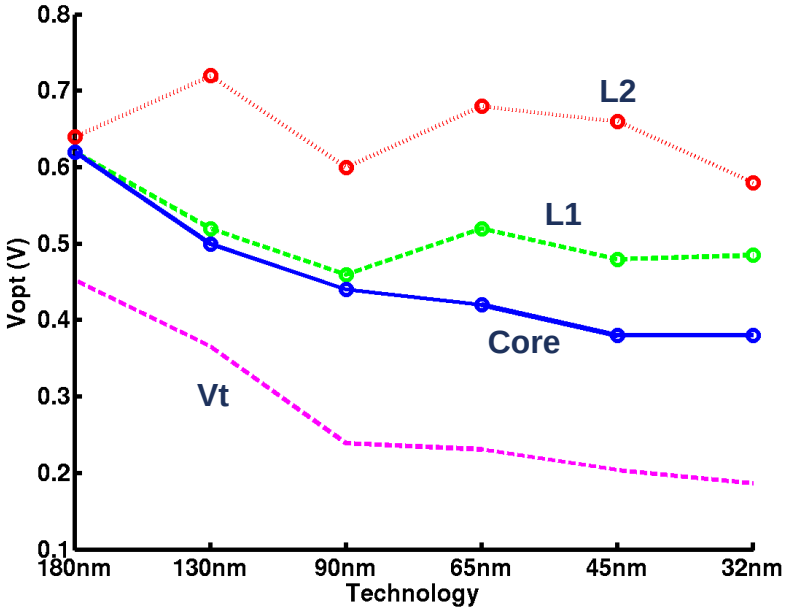


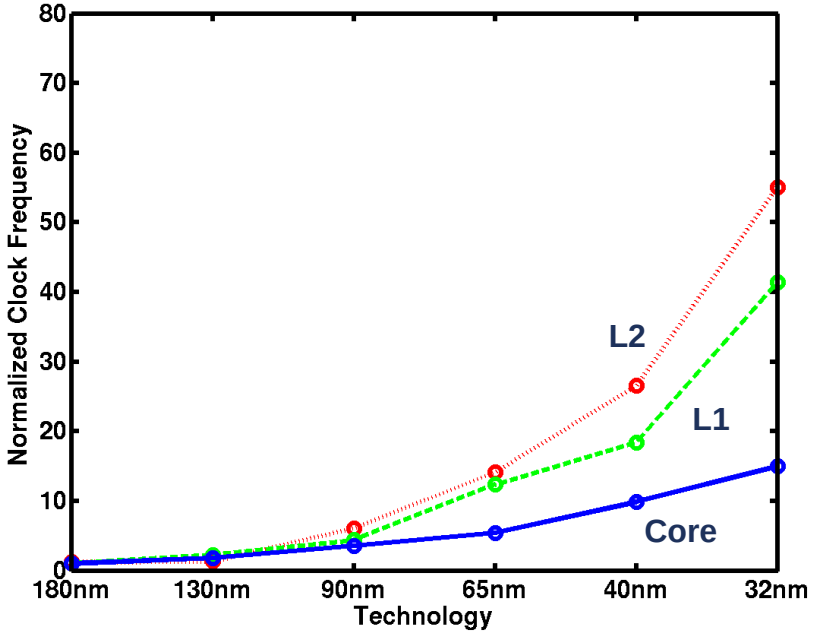
Near-Threshold Computing (NTC):

- *>60X power reduction*
- *6-8X energy reduction*
- *Enables 3D integration*



Architectural Impact of NTC

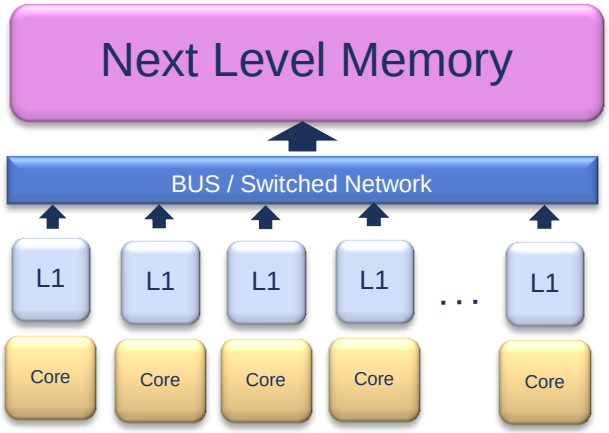


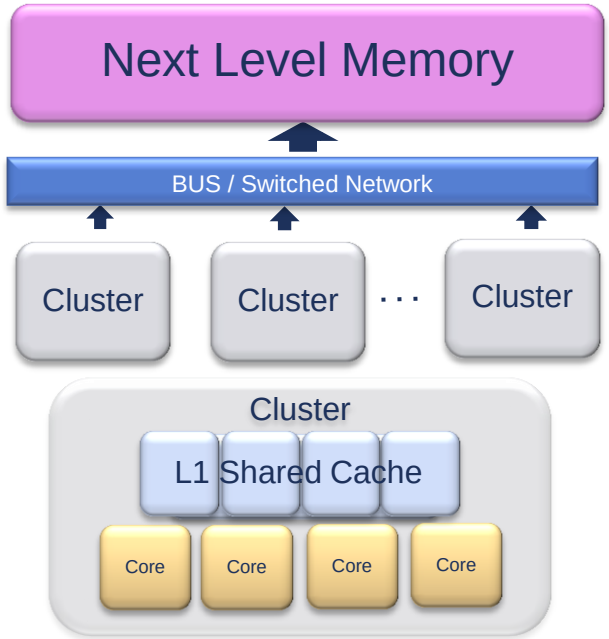


- Caches have higher Vopt and operating frequency
- Smaller activity rate when compared to core logic
- Leakage larger proportion of total power in caches
- New Architectures Possible

Proposed NTC Architecture

- SRAM is run at a higher V_{DD}
 - Caches operate faster than core
- Can introduce clustered architecture
 - Multiple cores share L1
 - Cores see private L1
 - L1 still provides single-cycle latency
- Advantages:
 - Less coherence/snoop traffic
 - Larger cache for processes that need it
- Drawbacks:
 - Core conflicts evicting L1 data
 - Not dominant in simulation
 - Longer interconnect
 - 3D addressable





Proposed Boosting Approach

Measured results for 130nm LP design

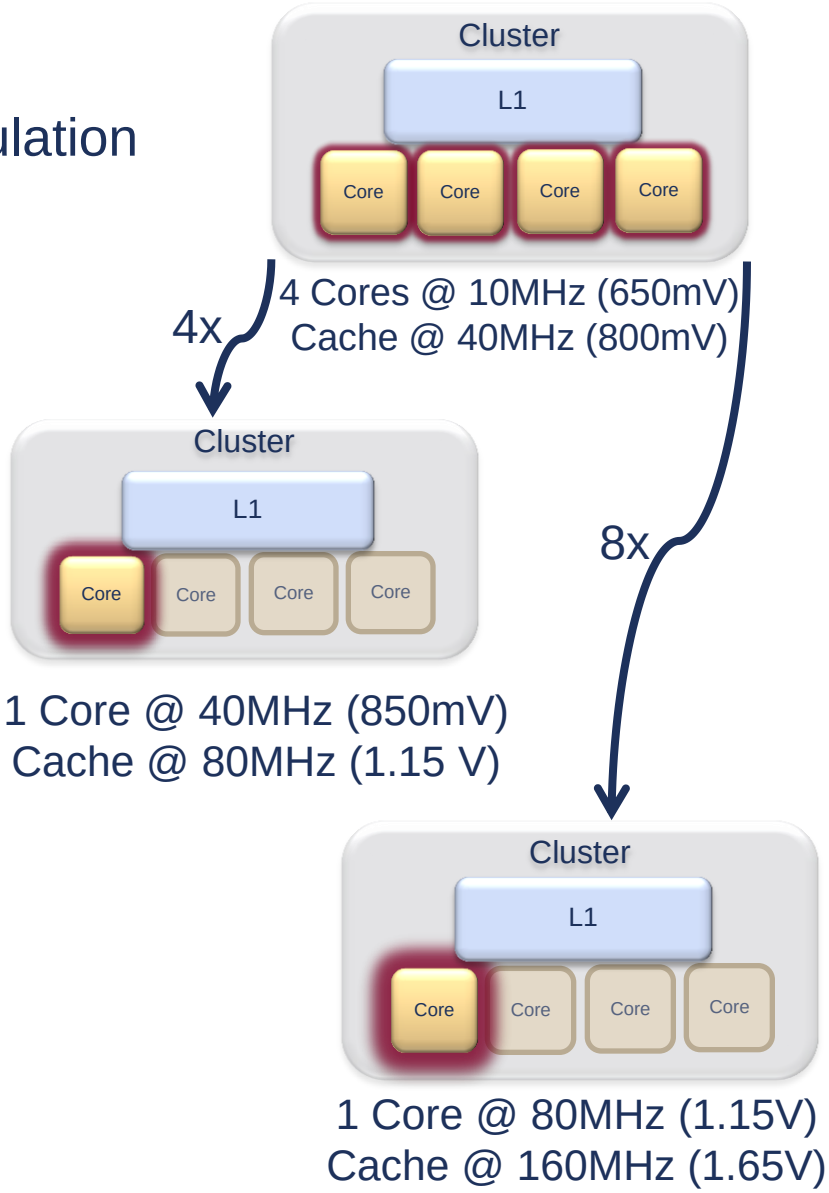
- 10MHz becomes ~110MHz in 32nm simulation
- 140 FO4 delay core

Baseline

- Cache runs 4x core frequency
- Pipelined cache

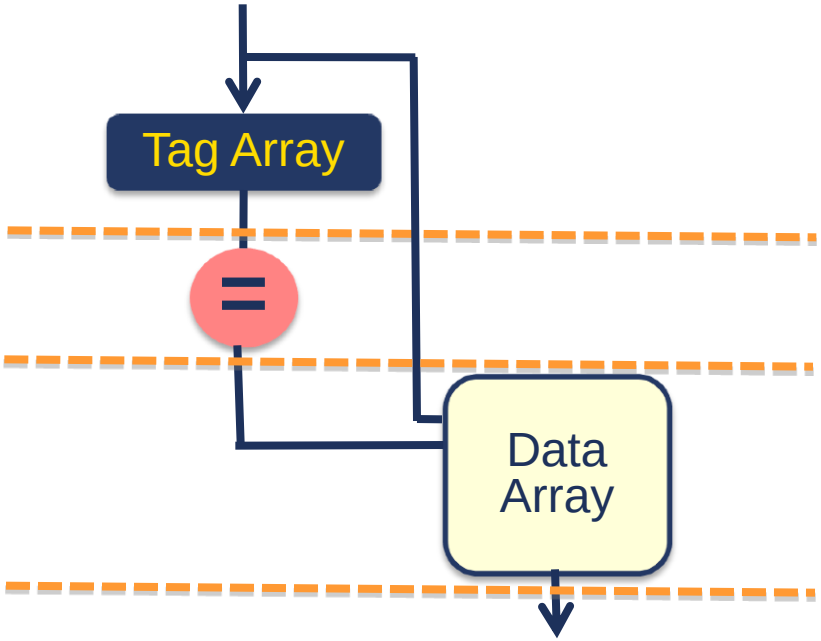
Better Single Thread Performance

- Turn some cores off, speed up the rest
- Cache de-pipelined
- Faster response time, *same* throughput
- Core sees larger cache
 - Faster cores needs larger caches

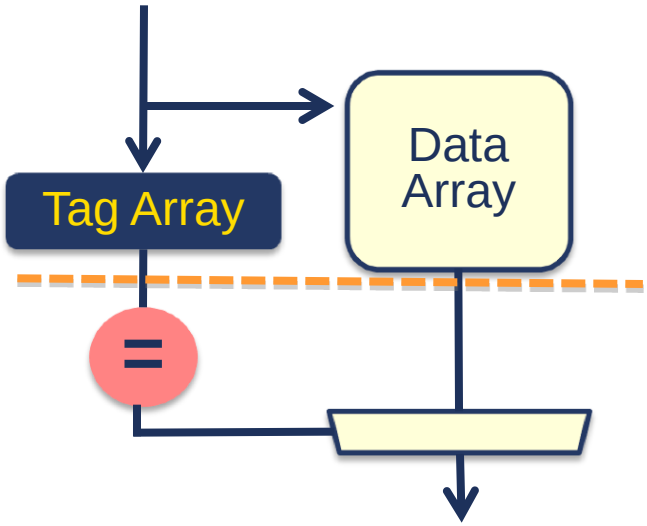


Cache Timing

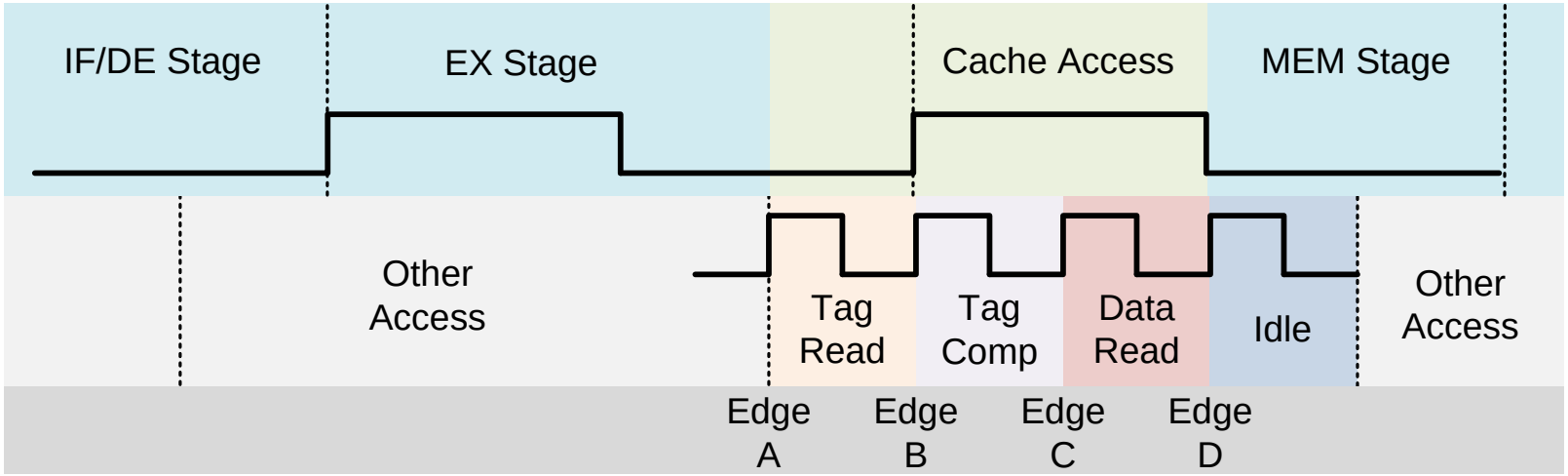
NTC Mode (3/4 Cores)
Low power
Tag arrays read first
0-1 data arrays accessed

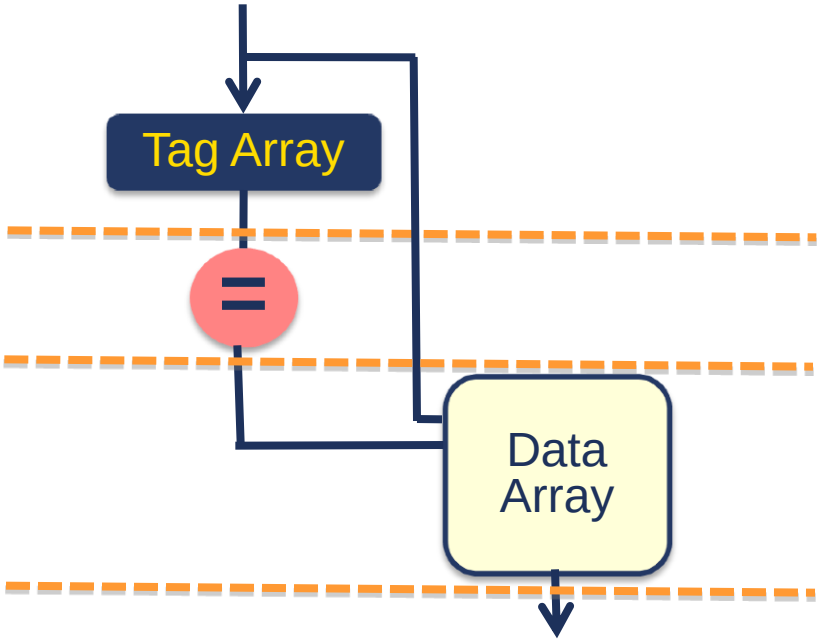


Boost Mode (1/2)
Low latency
Data and tags read in parallel
4 data arrays accessed



Cache Timing





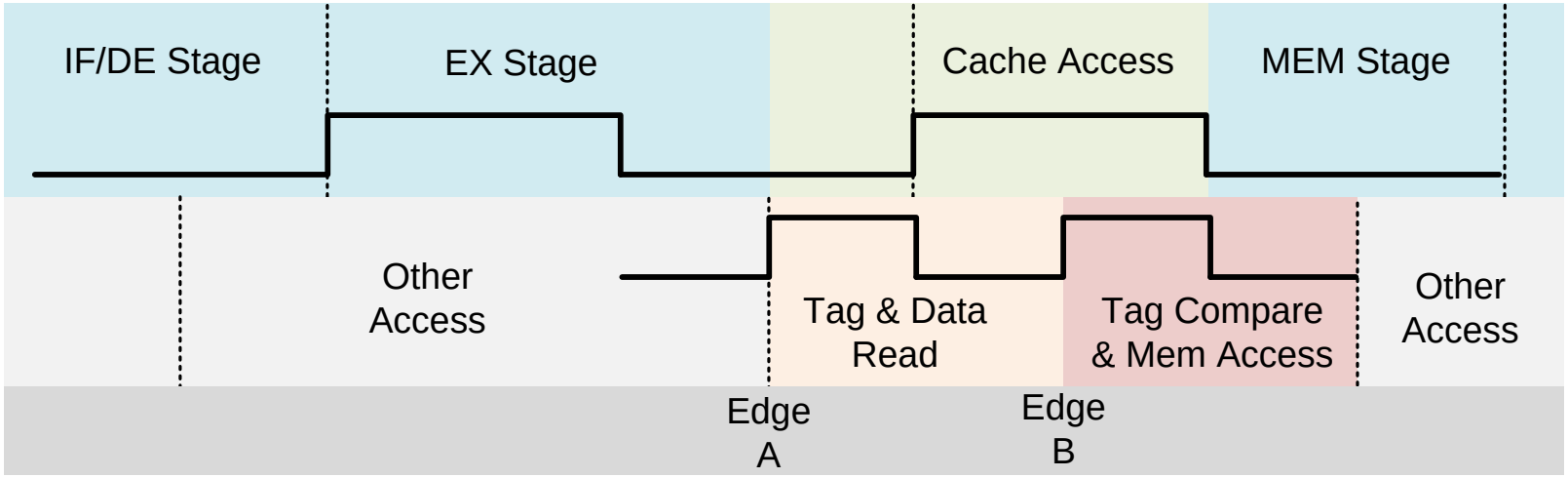
NTC Mode (3/4 Cores)

Low power

Tag arrays read first

0-1 data arrays accessed

Cache Timing

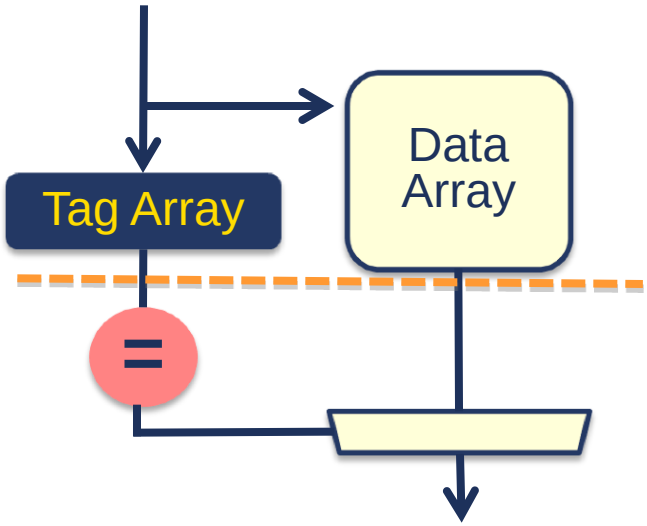


Boost Mode (1/2)

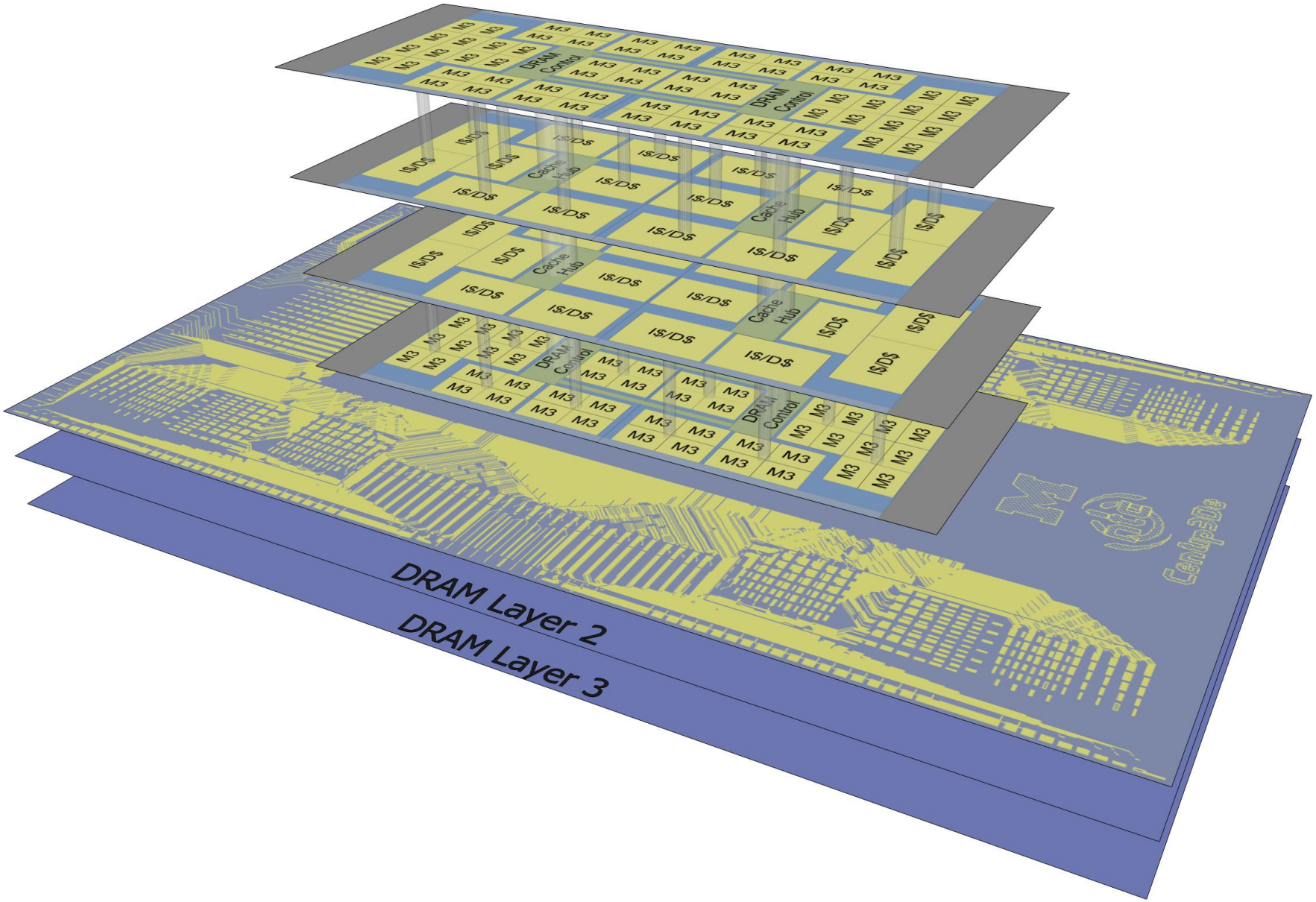
Low latency

Data and tags read in parallel

4 data arrays accessed

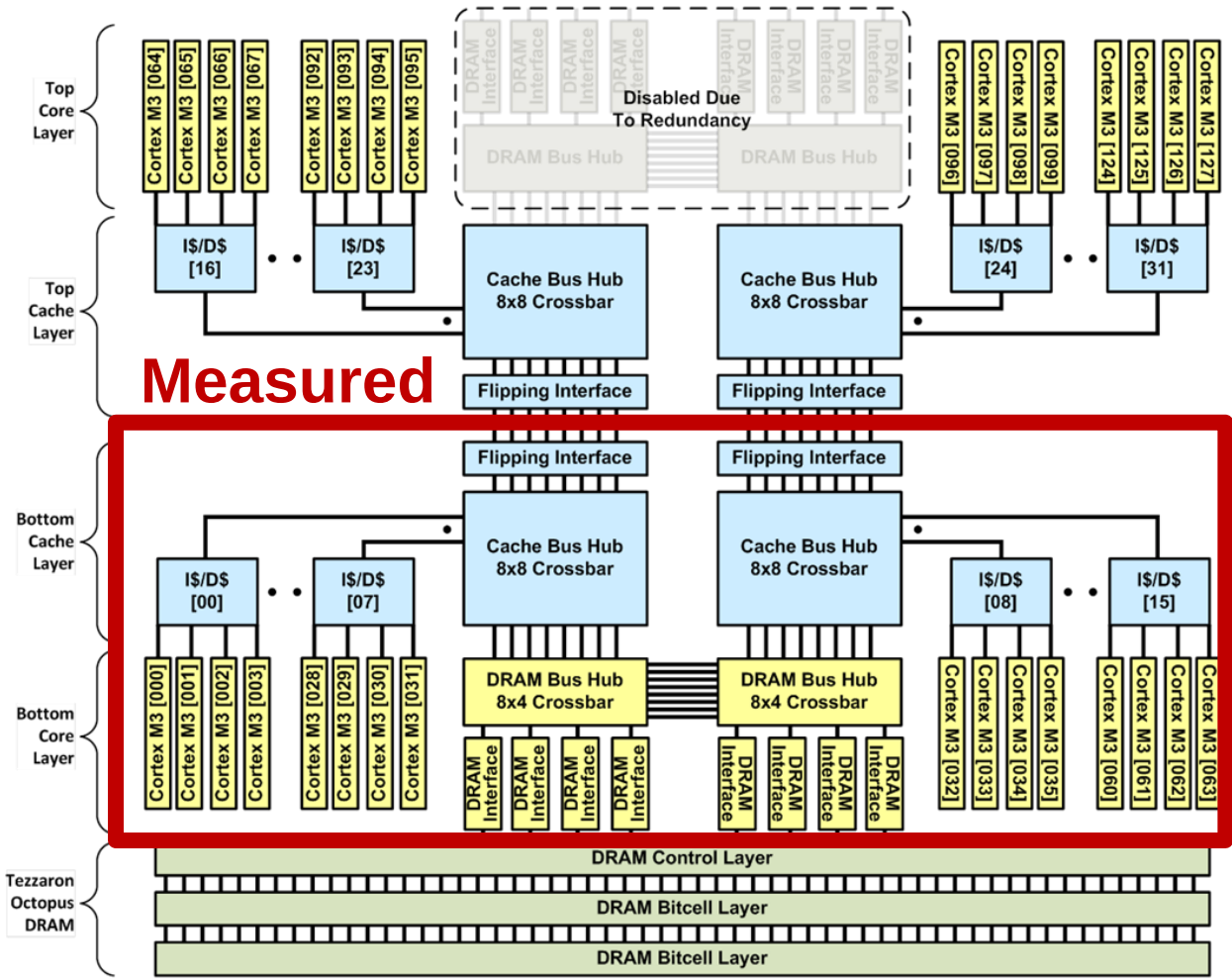


Centip3De System Overview



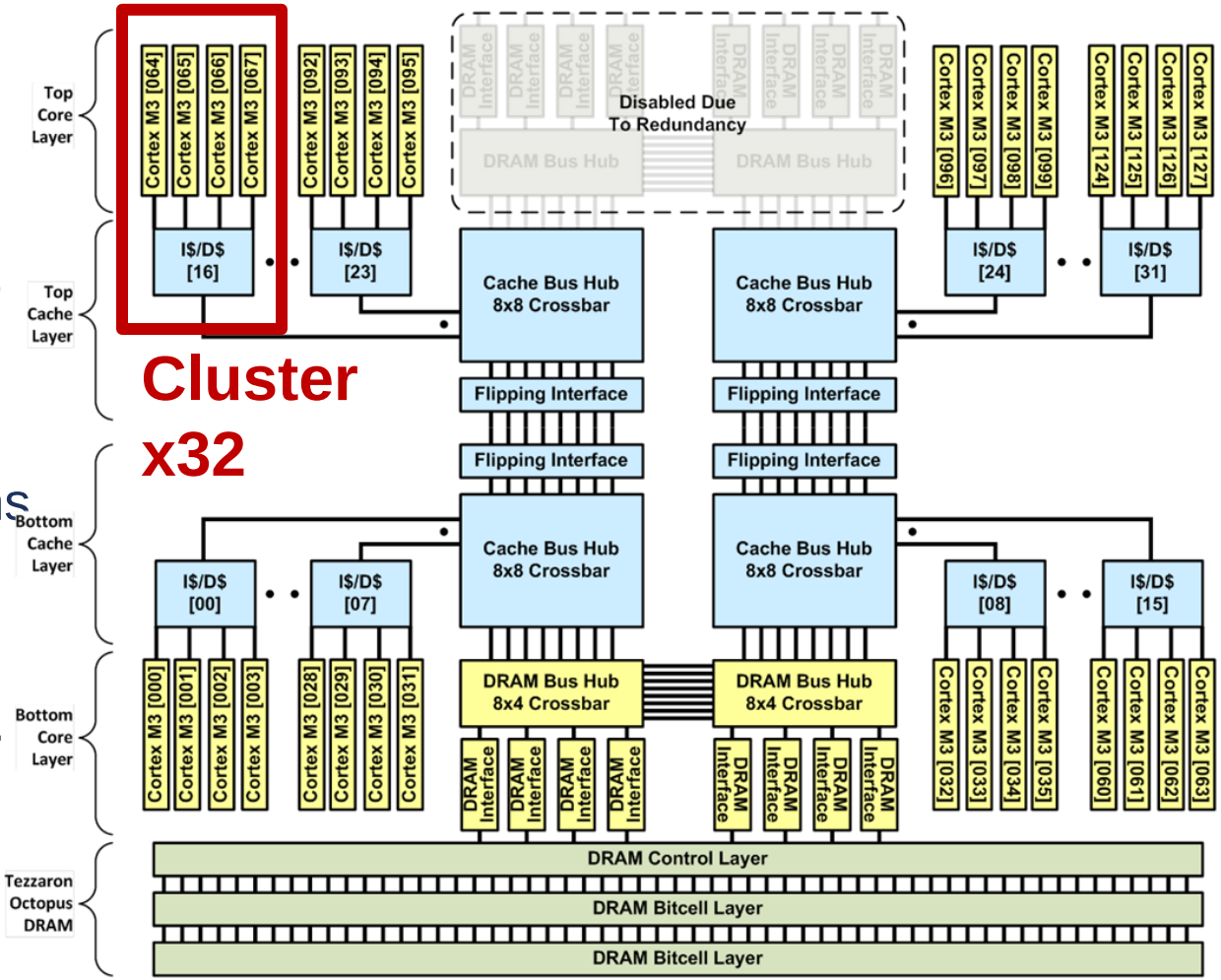
Centip3De System Overview

- 7-Layer NTC system
- 2-Layer system completed fabrication with measured results
- Full 7-layer system expected End of 2012



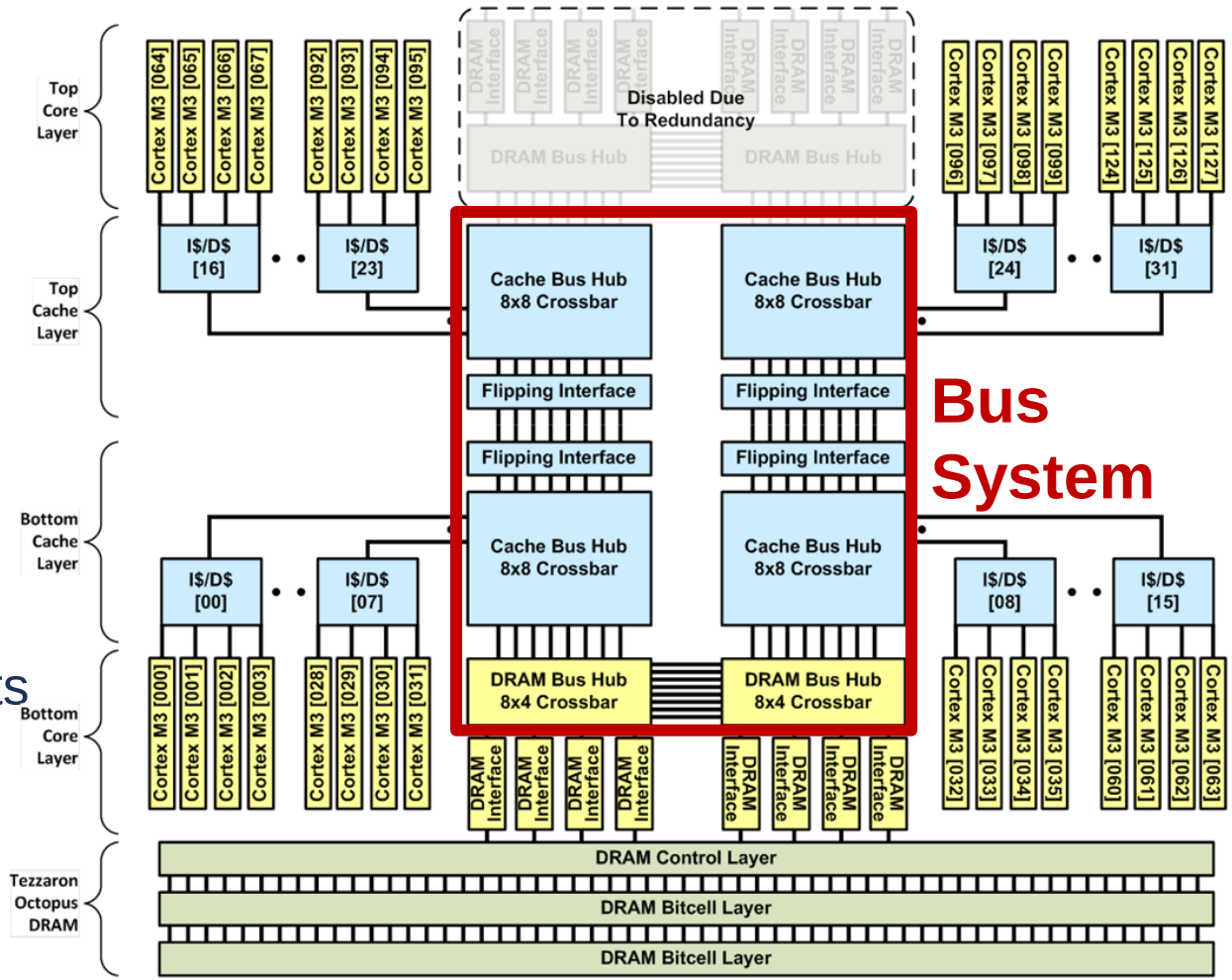
Centip3De System Overview

- Cluster architecture
 - 4 Cores/cluster
 - 1kB I$, 8kB D$
 - Local clock controller operates cores 90˚ Out-of-phase
 - 1591 F2F connections per cluster
- Organized into layer pairs (cache⇔core)
 - Minimizes routing
 - Up to two pairs
 - 16 clusters per pair
 - Cores have only vertical interconnections



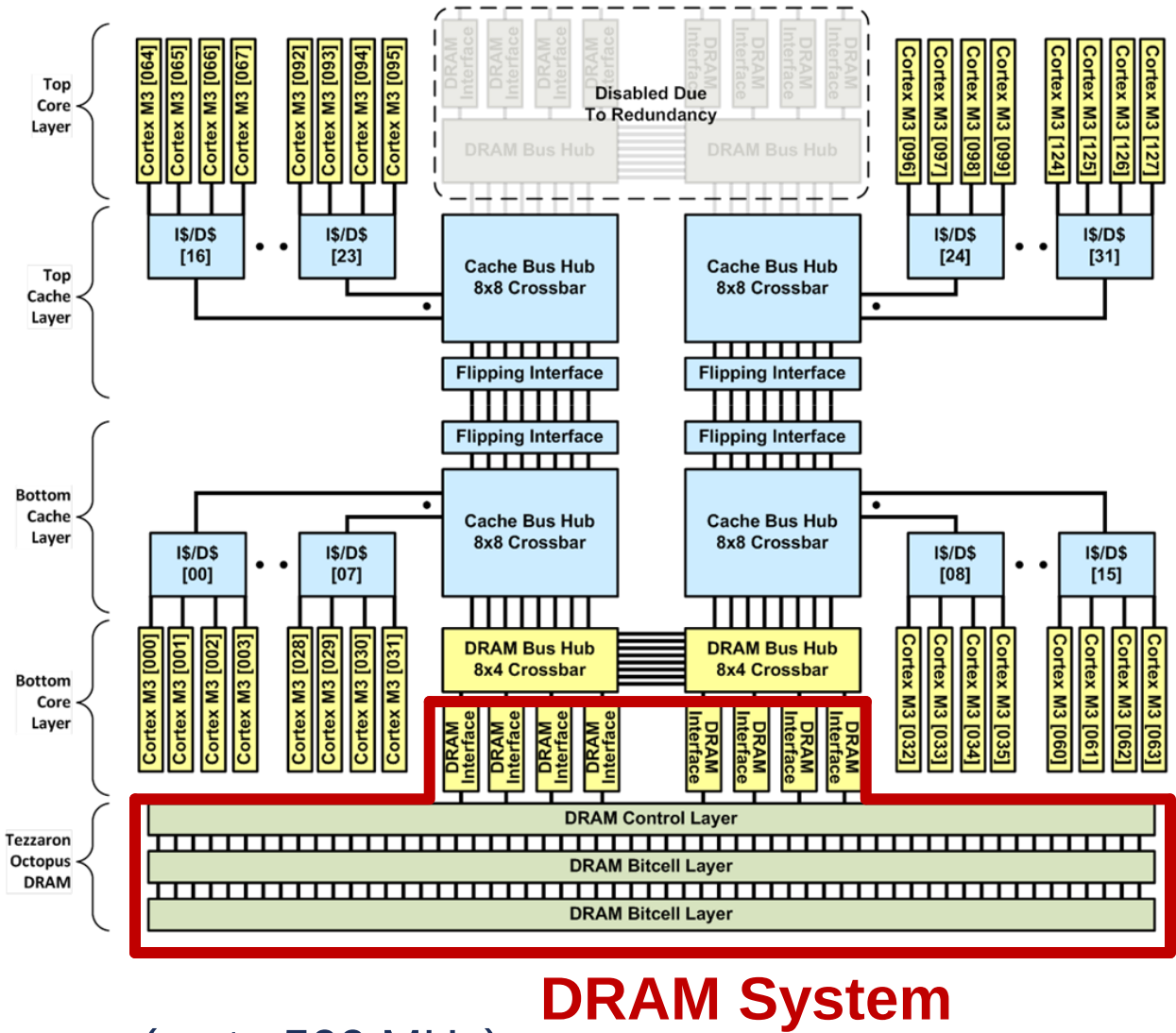
Centip3De System Overview

- Bus interconnect architecture
 - Up to 500 MHz
 - 9-11 cycle latency
 - 1-3 core cycles
- 8 lanes, each 128b
 - One per DRAM interface
 - Each cluster connects to all eight
 - 1024b total
- Vertically connected through all four layers
 - Flipping interface enables 128-core system

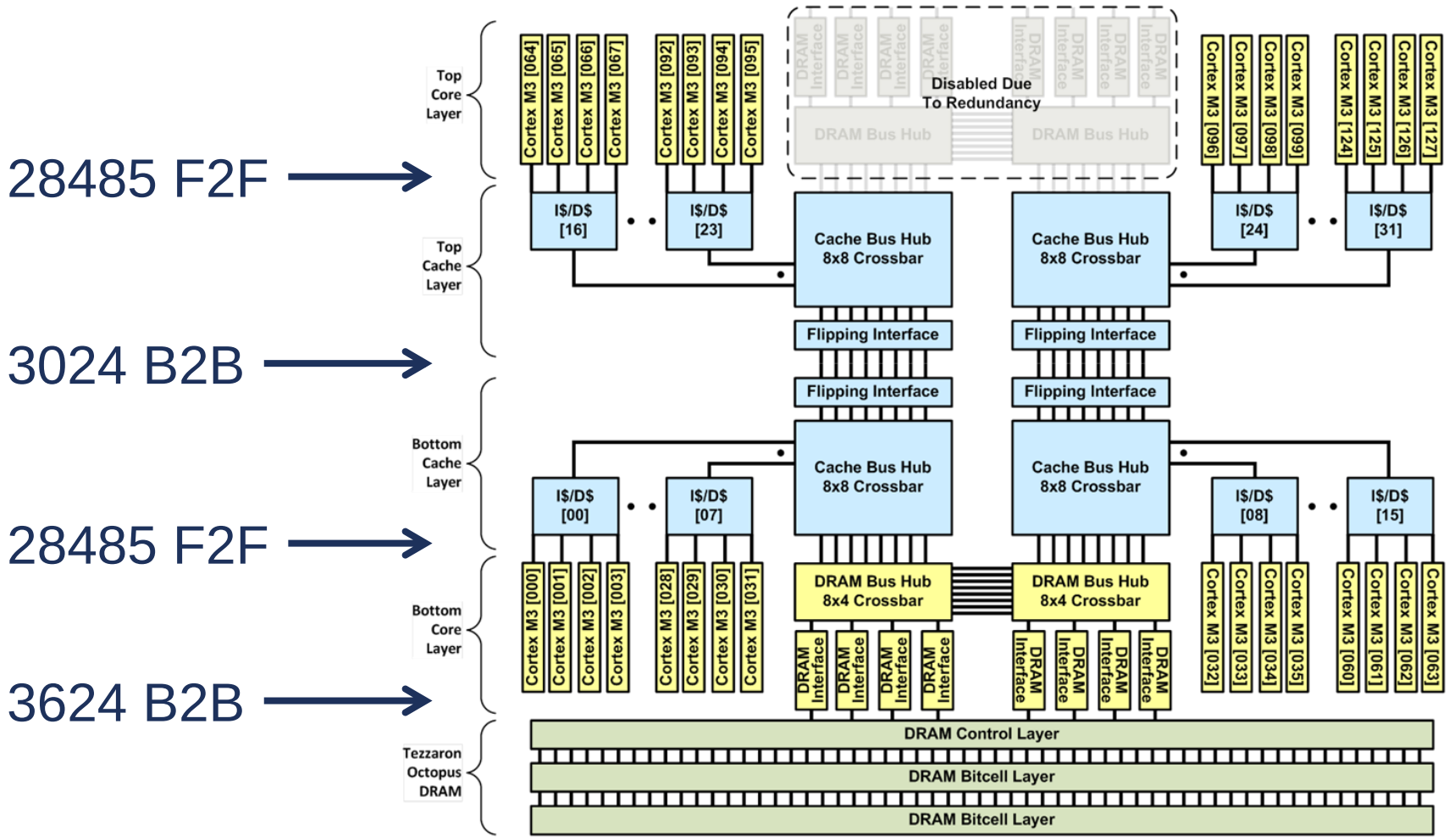


Centip3De System Overview

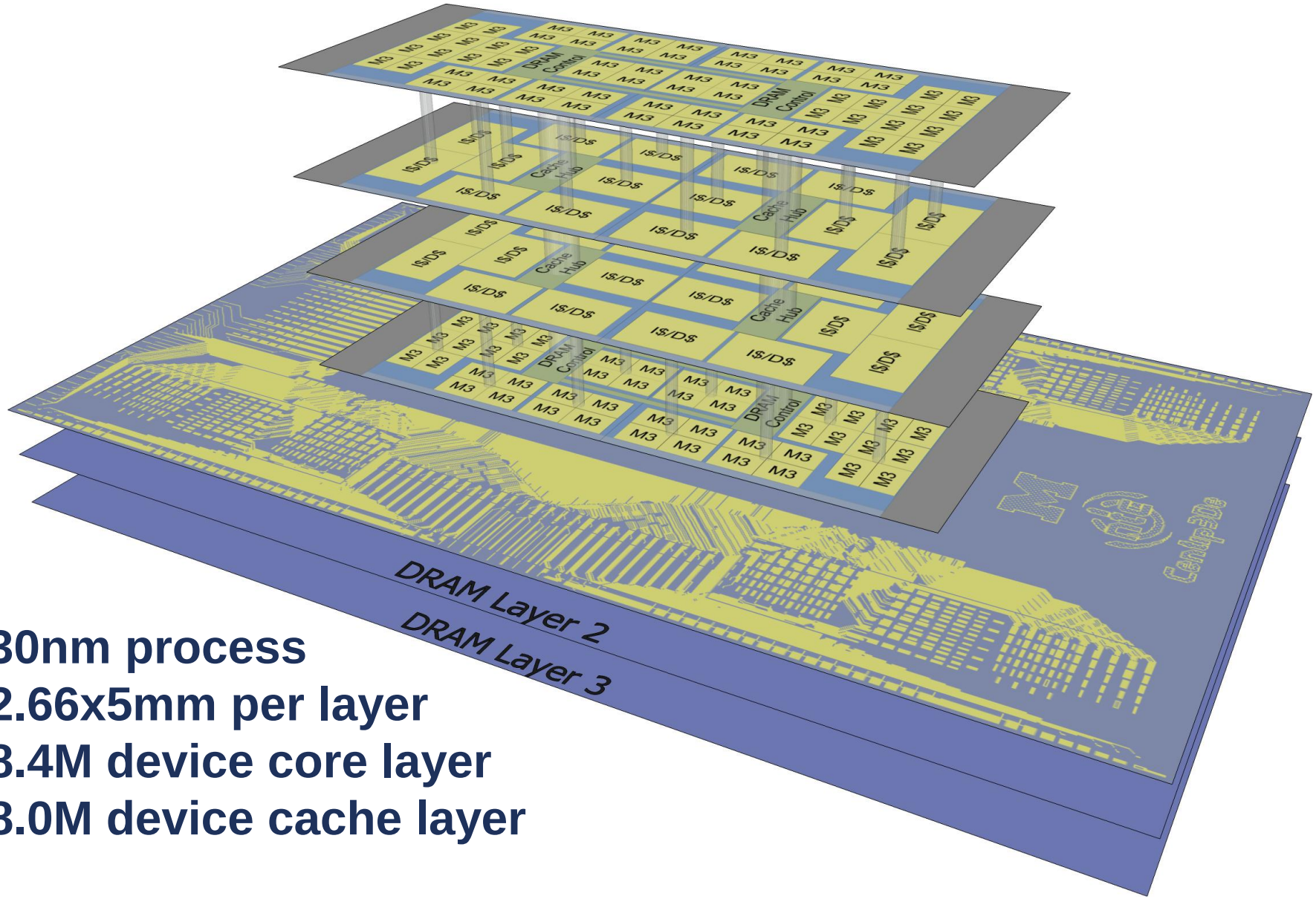
- 3D-Stacked DRAM
 - Tezzaron Octopus
- 1 control layer
 - 130nm CMOS
- 1 Gb bitcell layers
 - Up to two layers
 - DRAM process
- 8x 128b DDR2 interfaces
 - Operated at bus frequency (up to 500 MHz)



Centip3De System Overview

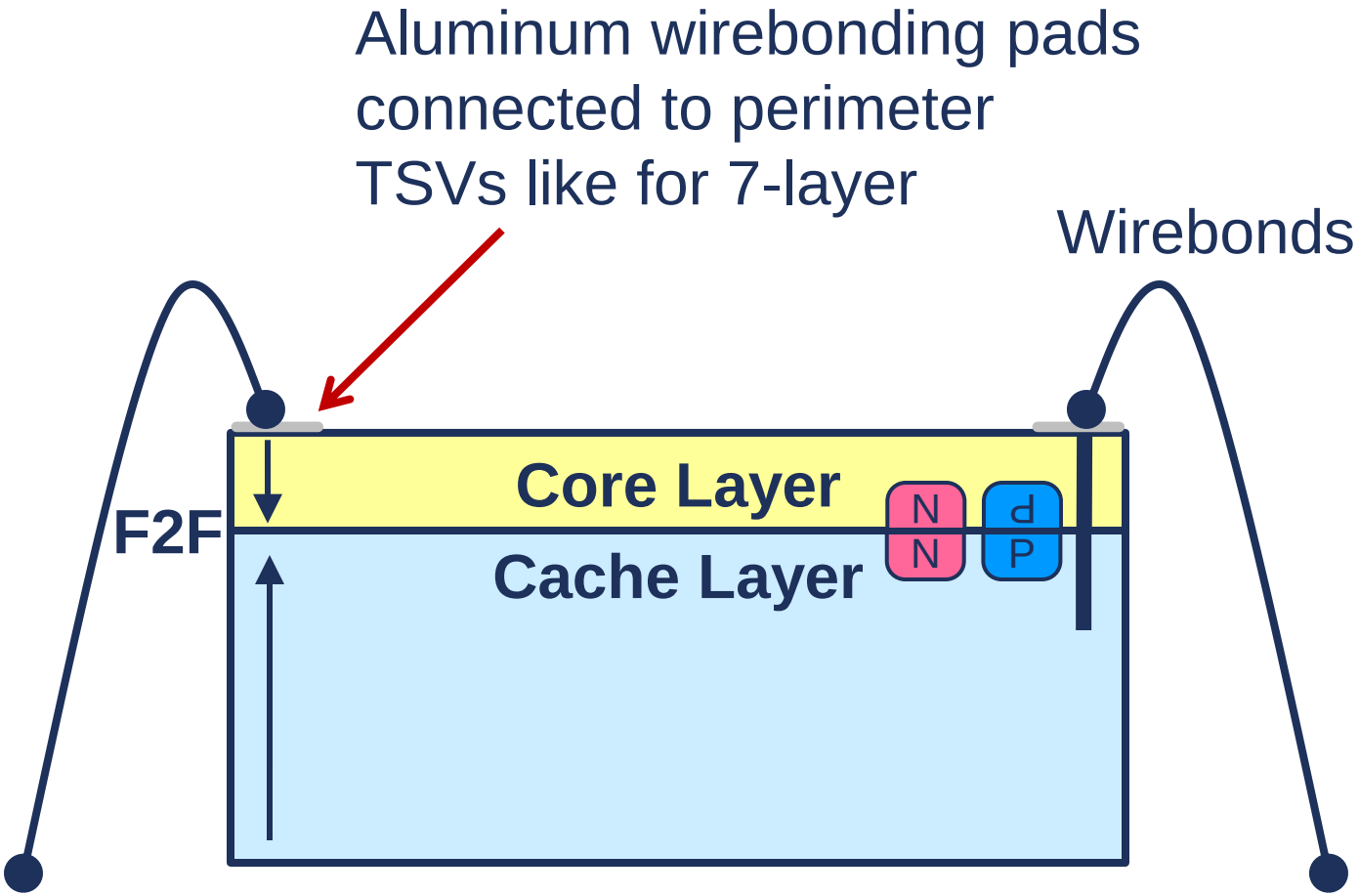


Centip3De System Overview



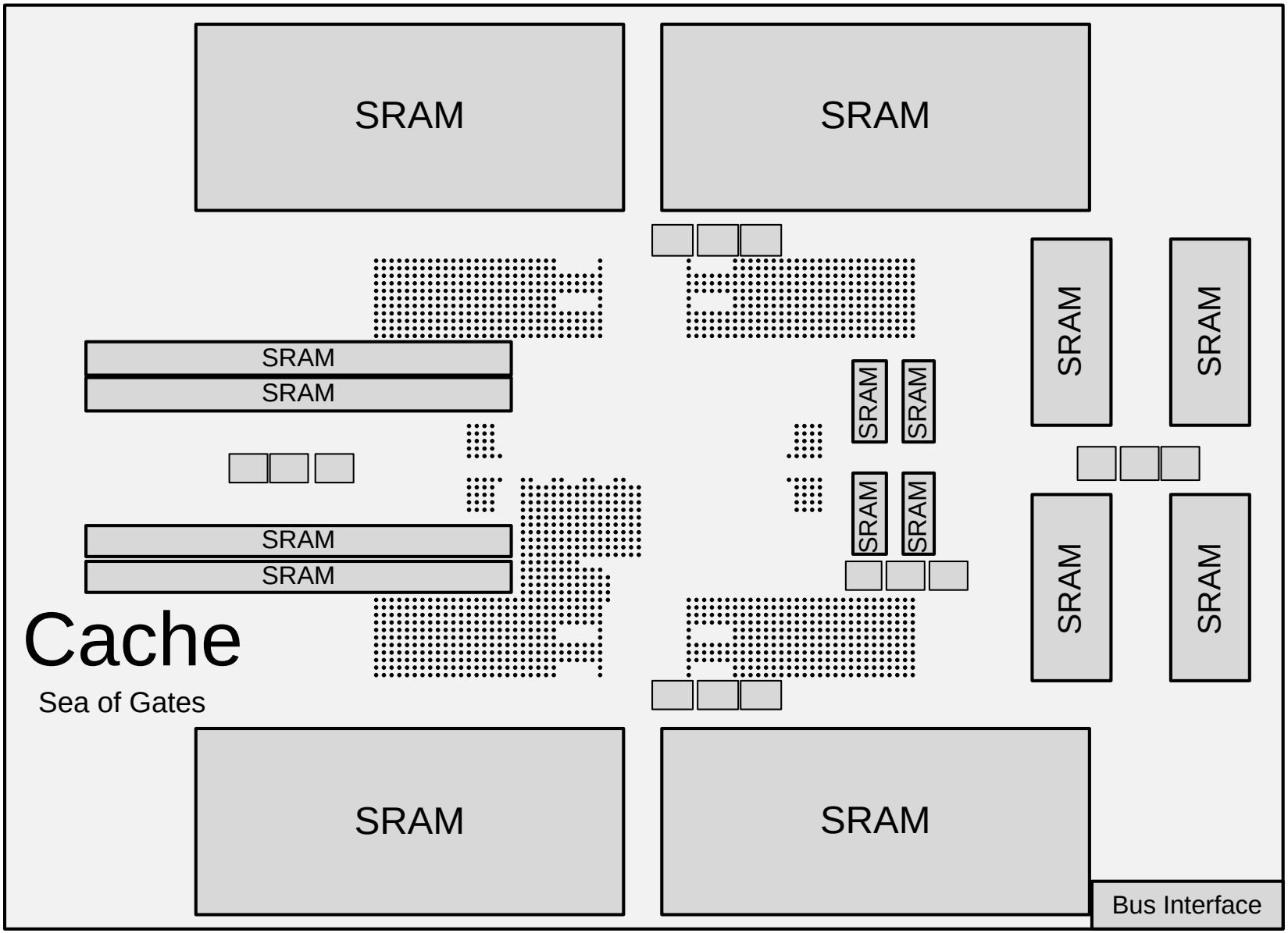
130nm process
12.66x5mm per layer
28.4M device core layer
18.0M device cache layer

2-Layer Stacking Process Evaluated

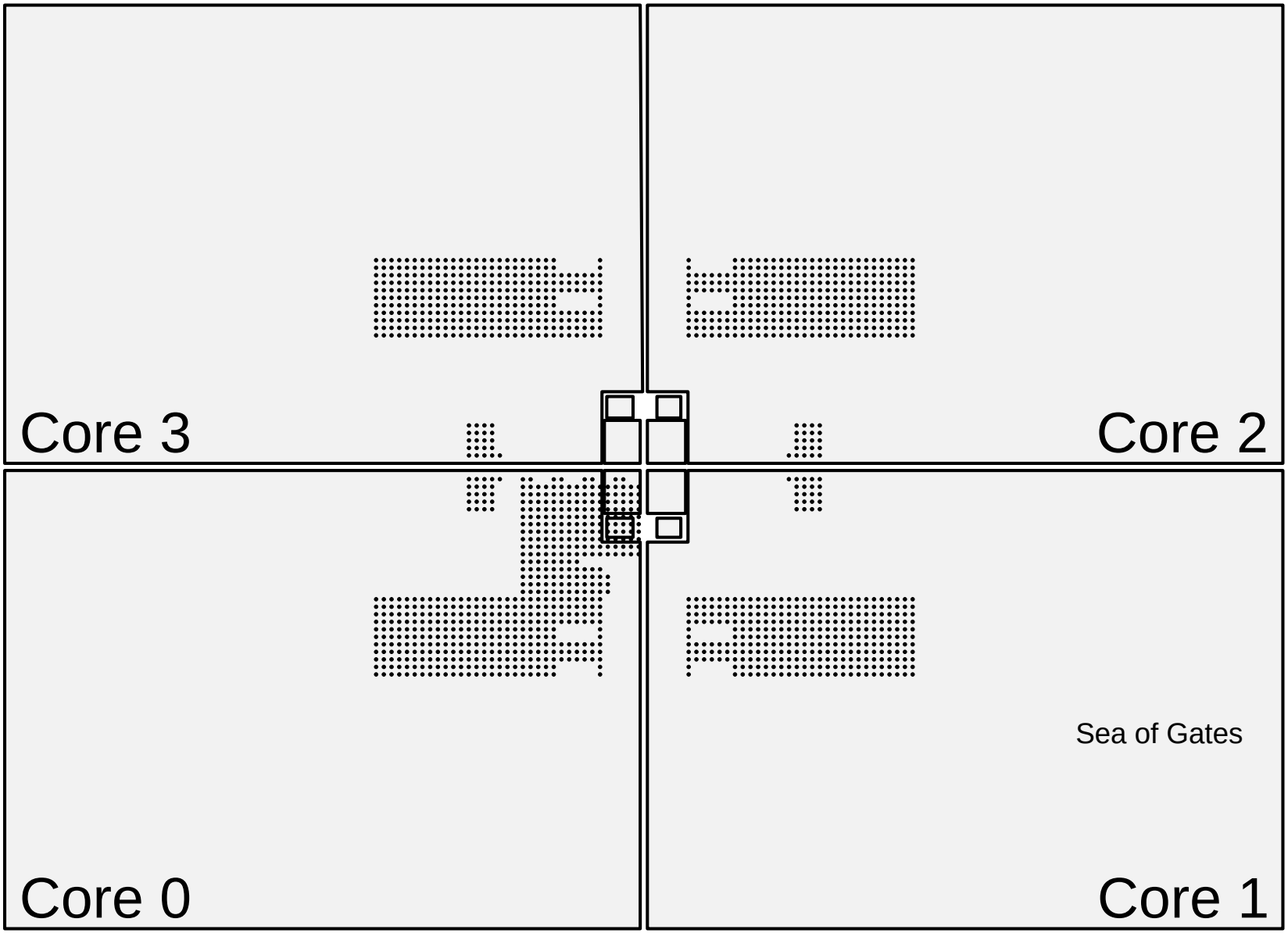


For the measured 2-layer system, aluminum wirebond pads were used instead

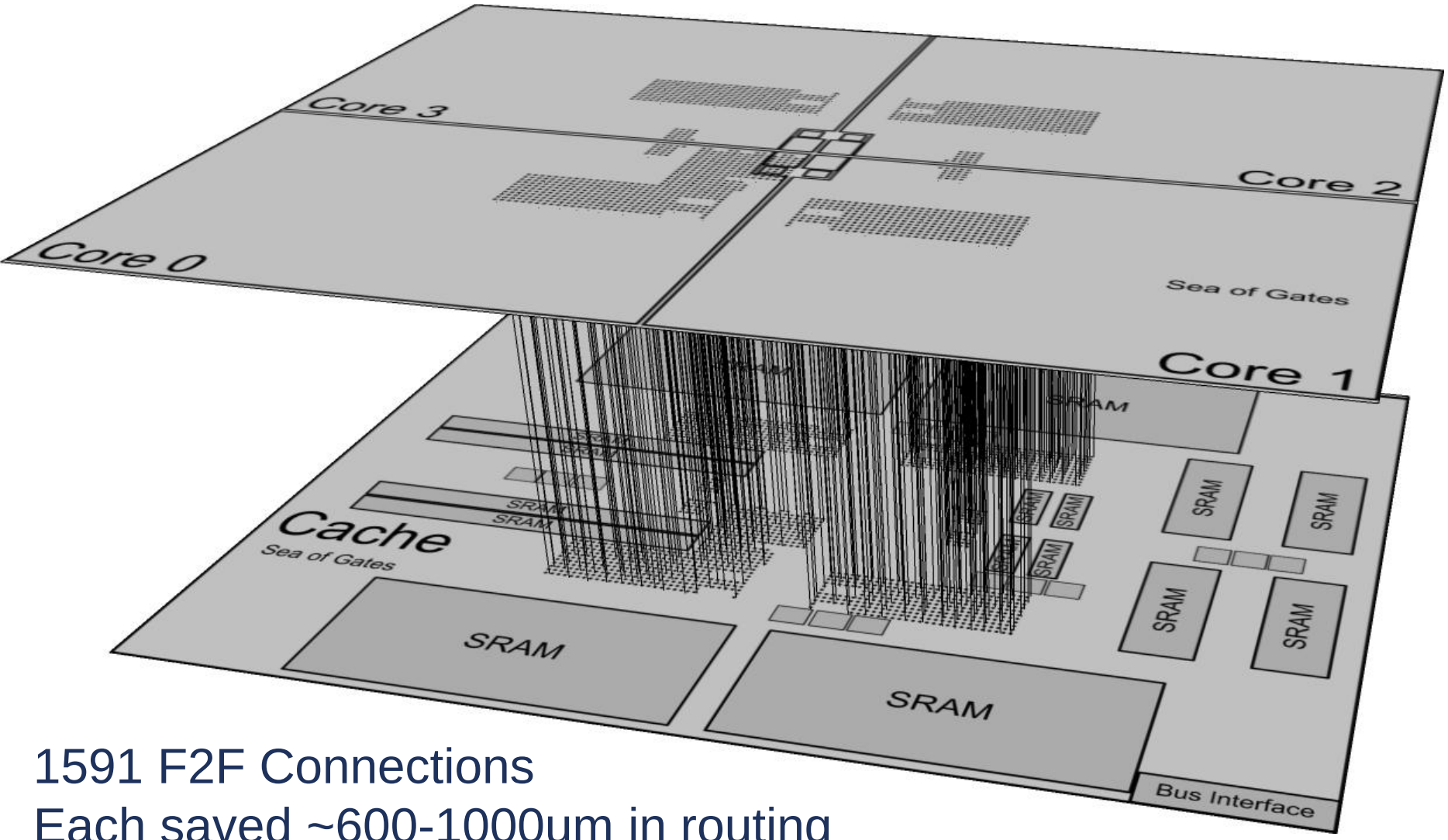
Cache 3D Connections



Core 3D Connections

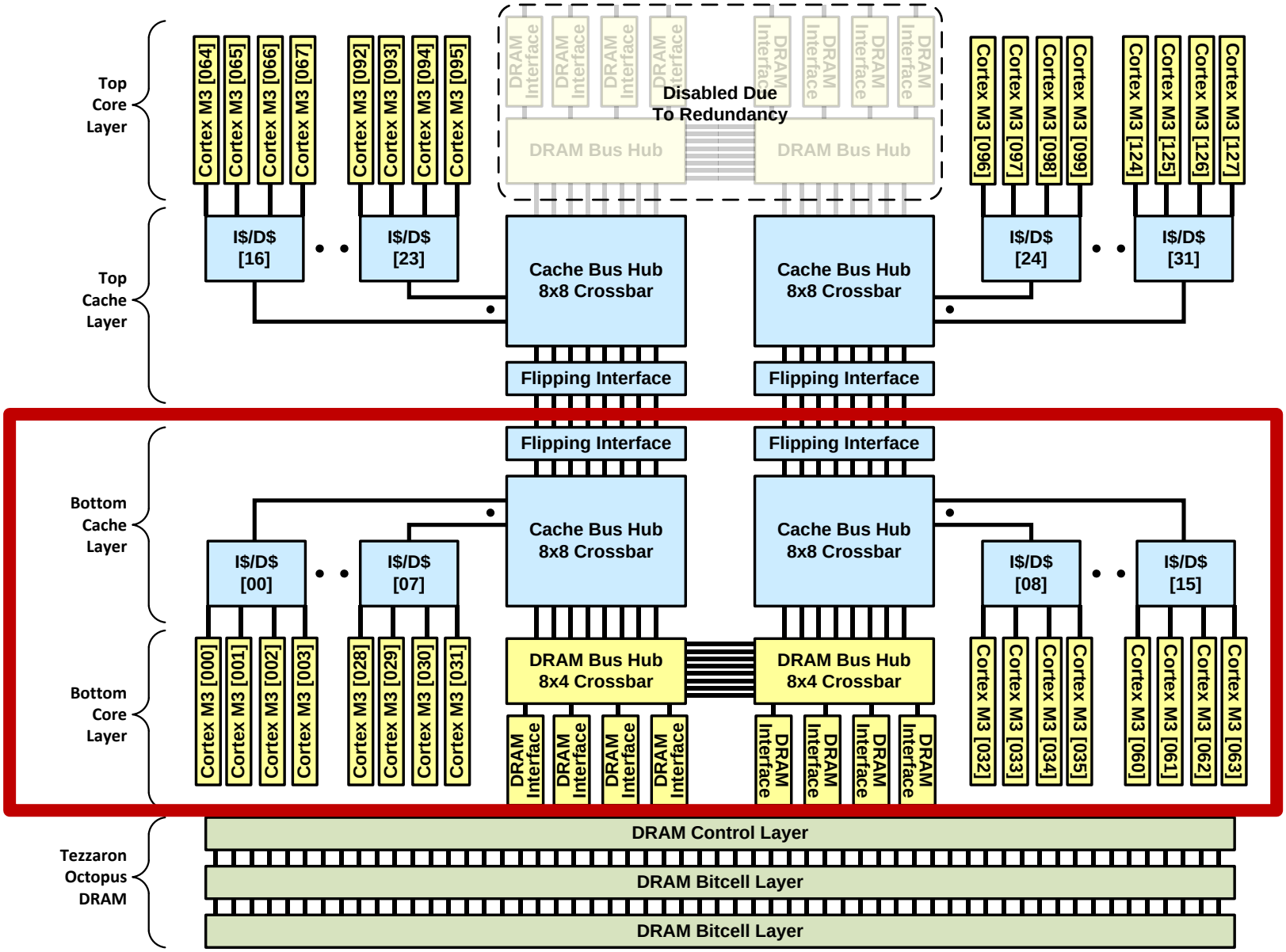


Cluster 3D Connections

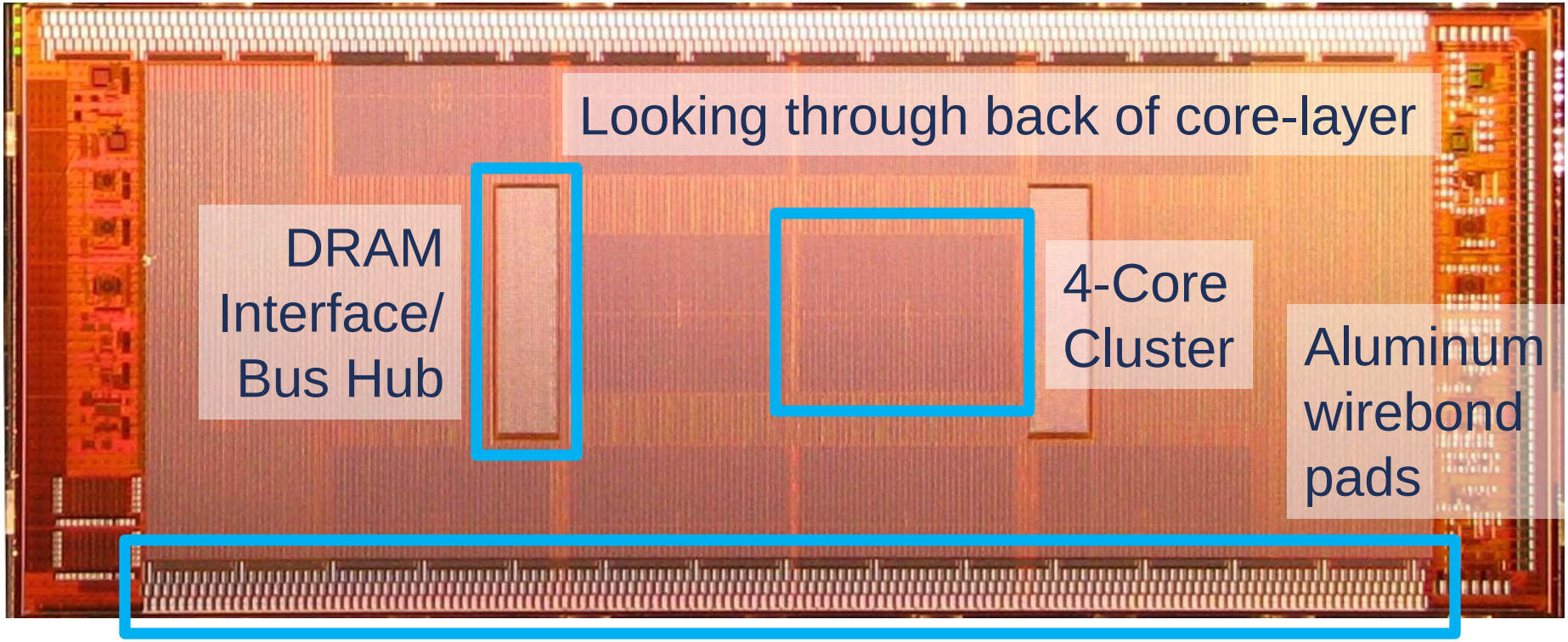


1591 F2F Connections
Each saved ~600-1000um in routing
Prevented wiring congestion around SRAMS

Silicon Results

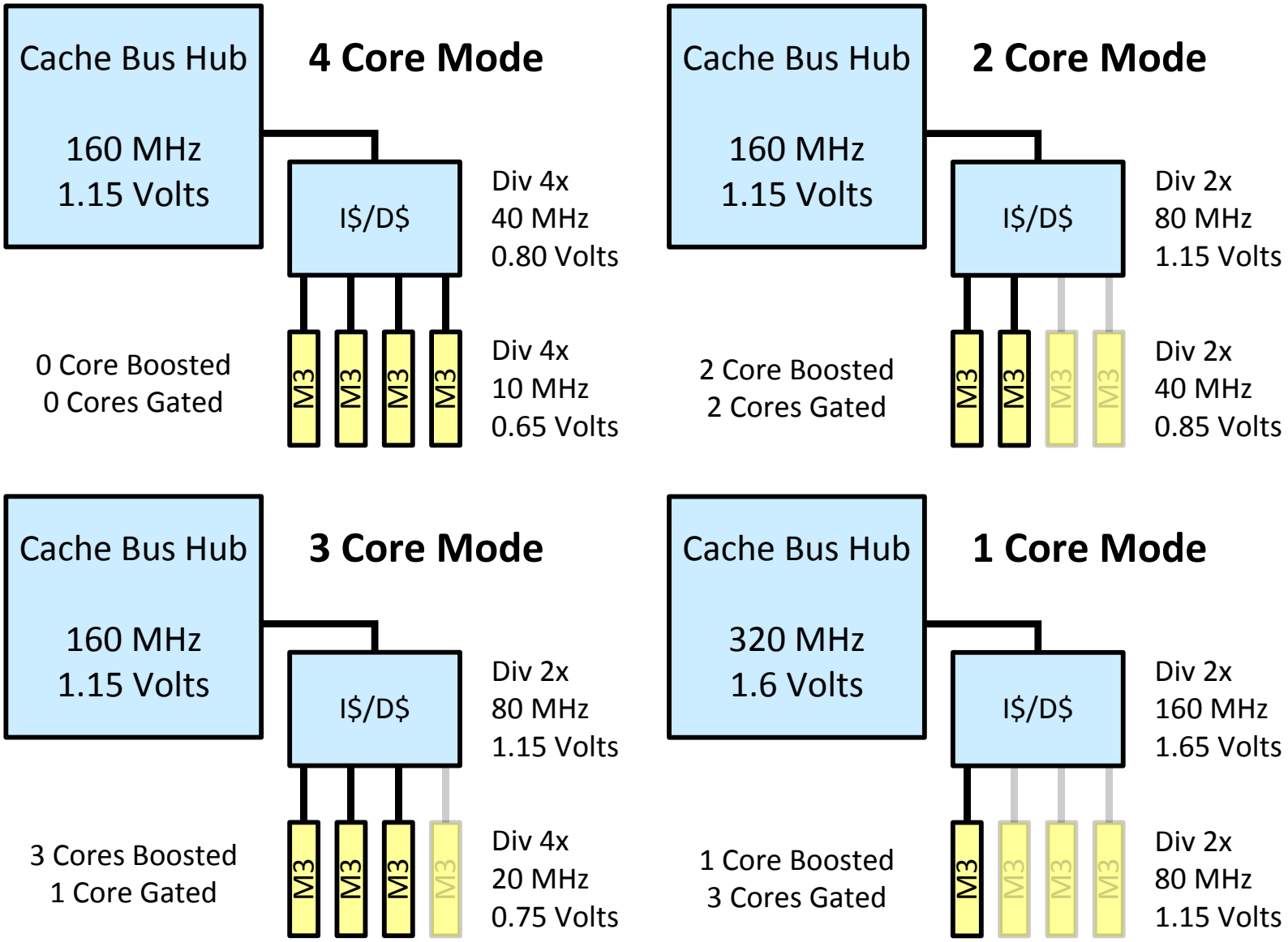


Die Shot



130nm process
12.66x5mm per layer
28.4M device core layer
18.0M device cache layer

System Configurations



Measured Results

Boosting a single cluster to 1-core mode requires disabling, or down-boosting other clusters

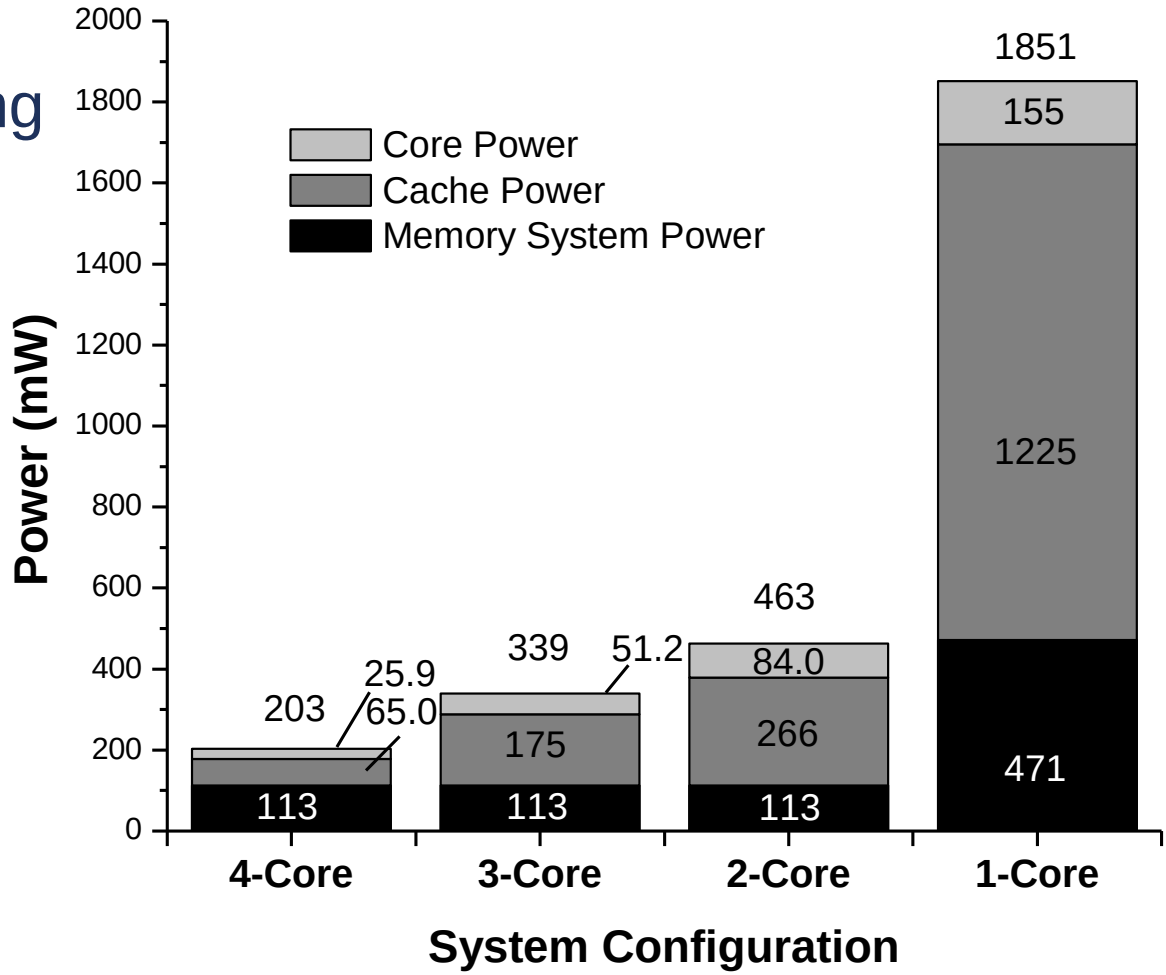
1-core cluster:

= 15x 4-core clusters

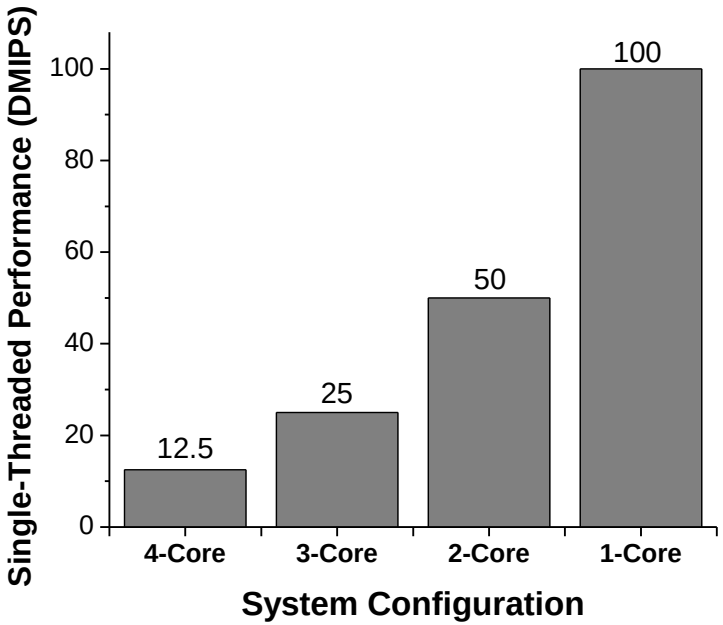
= 6x 3-core clusters

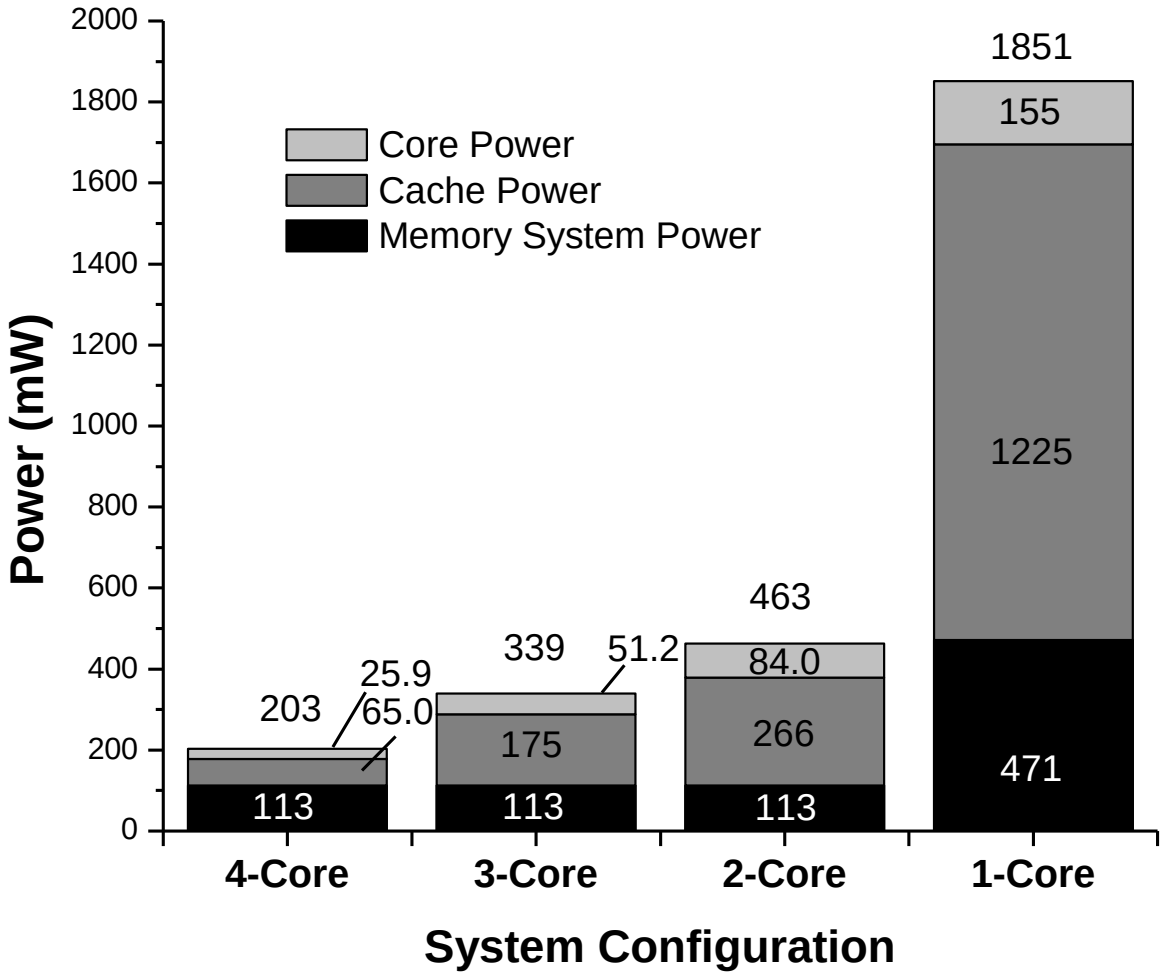
= 4.5x 2-core clusters

Baseline configuration depends on TDP and processing needs

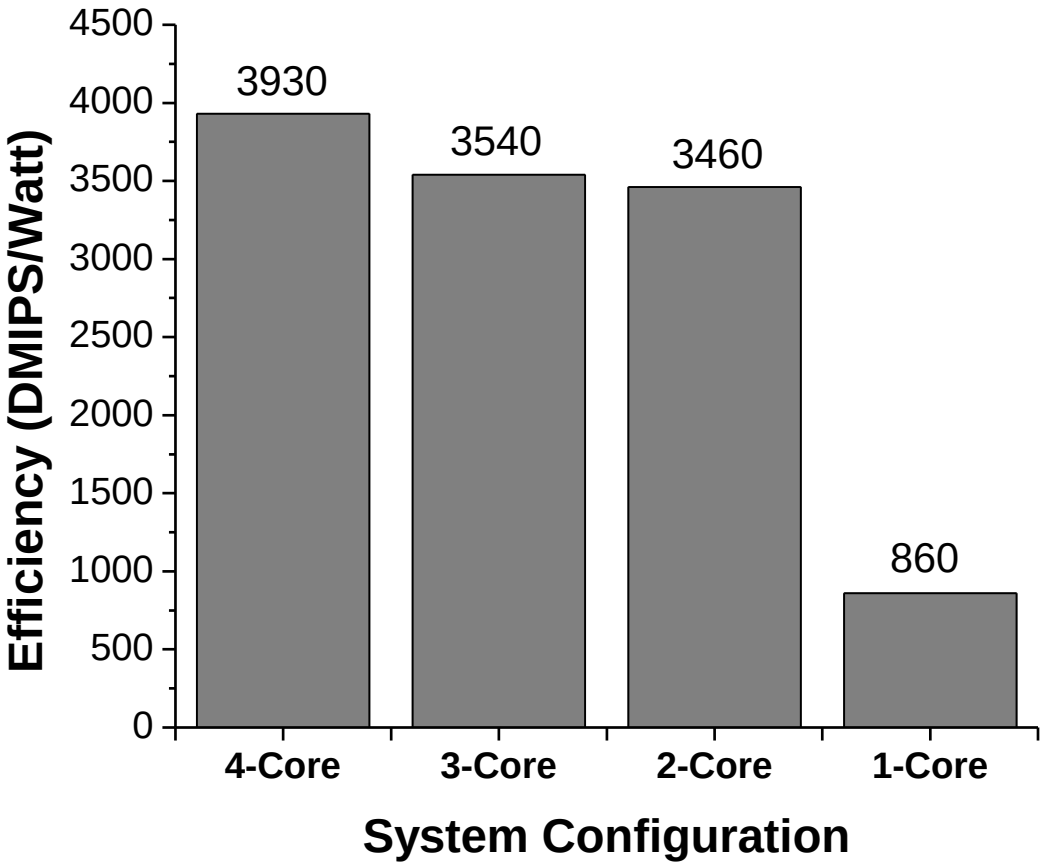


Measured Results





Measured Results



Measured Results:
Centip3De – 3,930 (130nm)

Industry Comparison:
ARM A9 – 8,000 (40nm) [1]

Estimated Results:
Centip3De – 18,500 (45nm)

[1] http://arm.com/products/processors/cortex-a/cortex-a9.php, ARM Ltd, 2011.

Conclusion

- Near threshold computing (NTC)
 - Need low power solutions to maintain TDP
 - Achieves 10x energy efficiency => 10x more computation to give TDP
 - Offers optimum balance between performance and energy
 - Allows boosting for single threaded performance (Amdahl's law)

- Large scale 3D CMP demonstrated
 - 64 cores currently
 - 128 cores + DRAM in the future
 - 3D design shown to be feasible

- This work was funded and organized with the help of DARPA, Tezzaron, ARM, and the National Science Foundation



FPGAs with 28Gb/s Transceivers Built with Heterogeneous Stacked-Silicon Interconnects

Ephrem Wu and Suresh Ramalingam

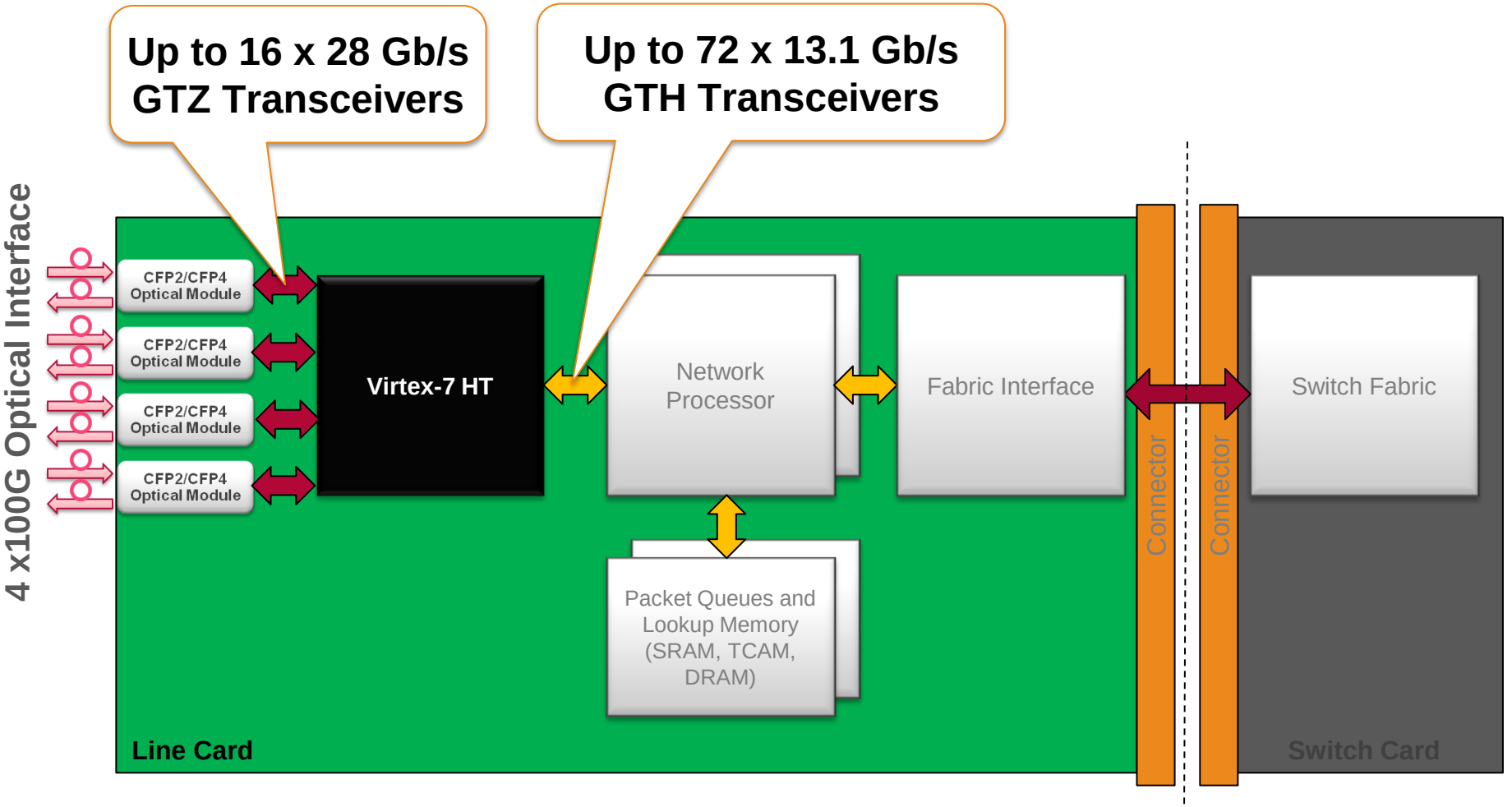
© Copyright 2012 Xilinx

Outline

1. Key Application
2. Heterogeneous Stacked-Silicon FPGA Family
3. Stacked-Silicon Packaging
4. Two Types of Stacked-Silicon Interconnects

© Copyright 2012 Xilinx

400Gb/s Line Card Application

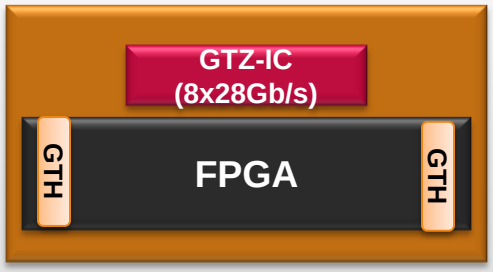


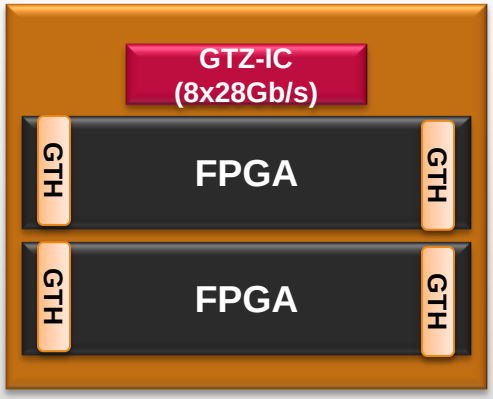
© Copyright 2012 Xilinx

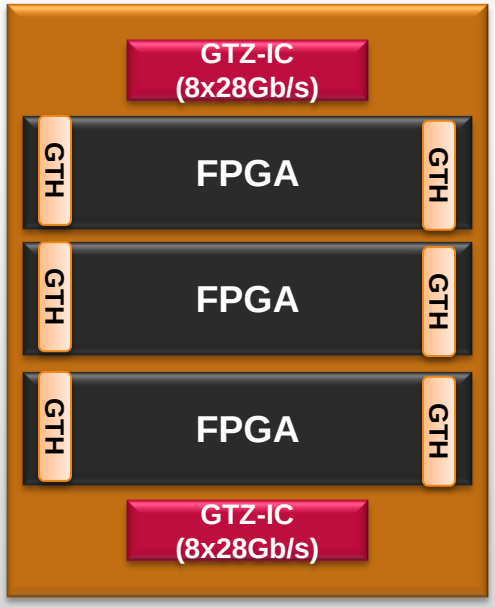
XILINX ALL PROGRAMMABLE

Heterogeneous Stacked-Silicon FPGAs

Interposer Floorplans







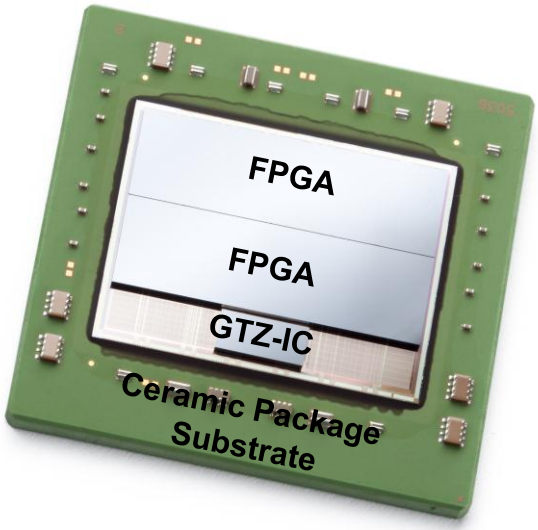
	XC7VH290T	XC7VH580T	XC7VH870T
Network	2 x 100 Gb/s	2 x 100 Gb/s	4 x 100 Gb/s
GTZ (28G)	8	8	16
GTH (13G)	24	48	72

© Copyright 2012 Xilinx

XILINX ALL PROGRAMMABLE

XC7VH580T Under the Hood

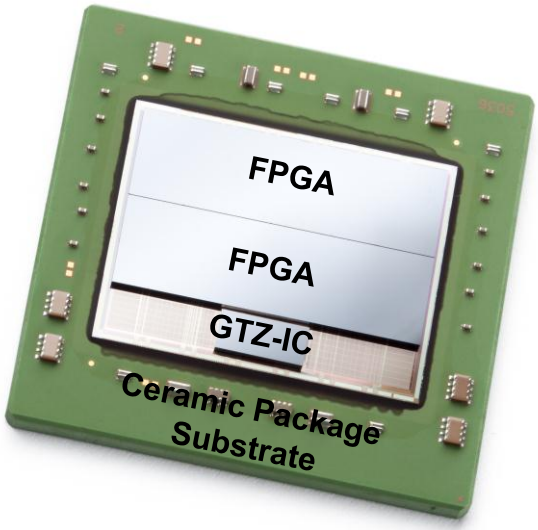
Industry's First Heterogeneous FPGA

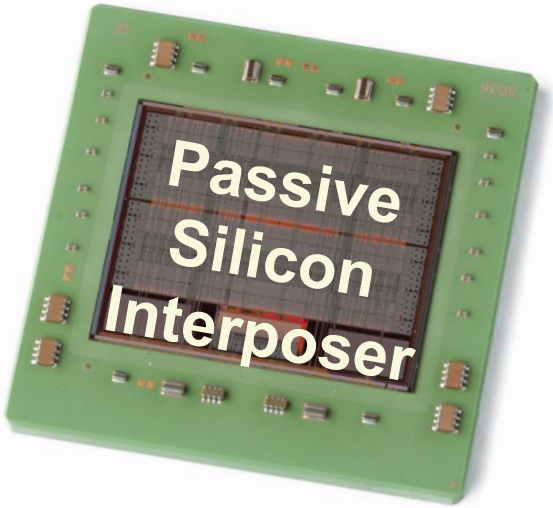


© Copyright 2012 Xilinx

XC7VH580T Under the Hood

Industry's First Heterogeneous FPGA



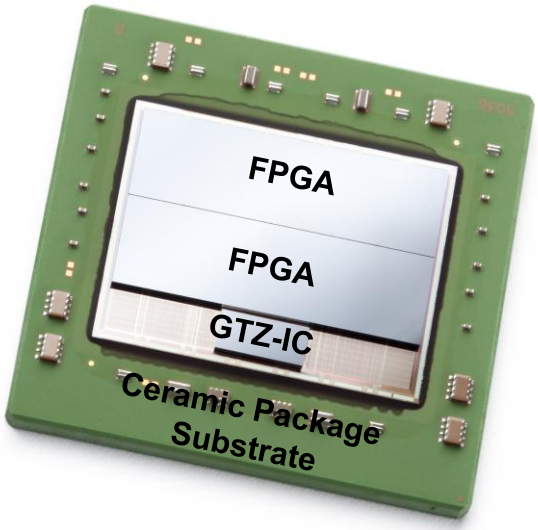


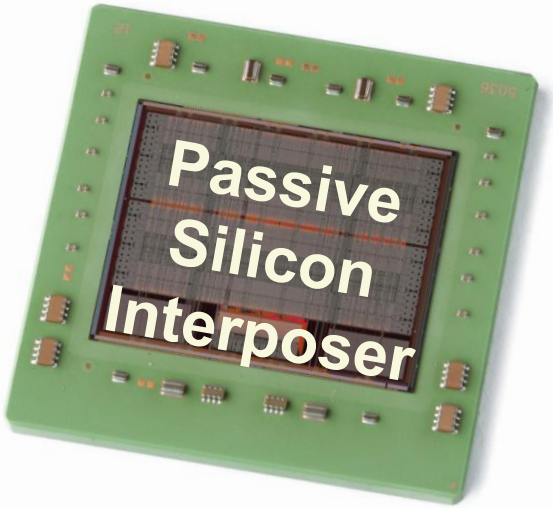
© Copyright 2012 Xilinx

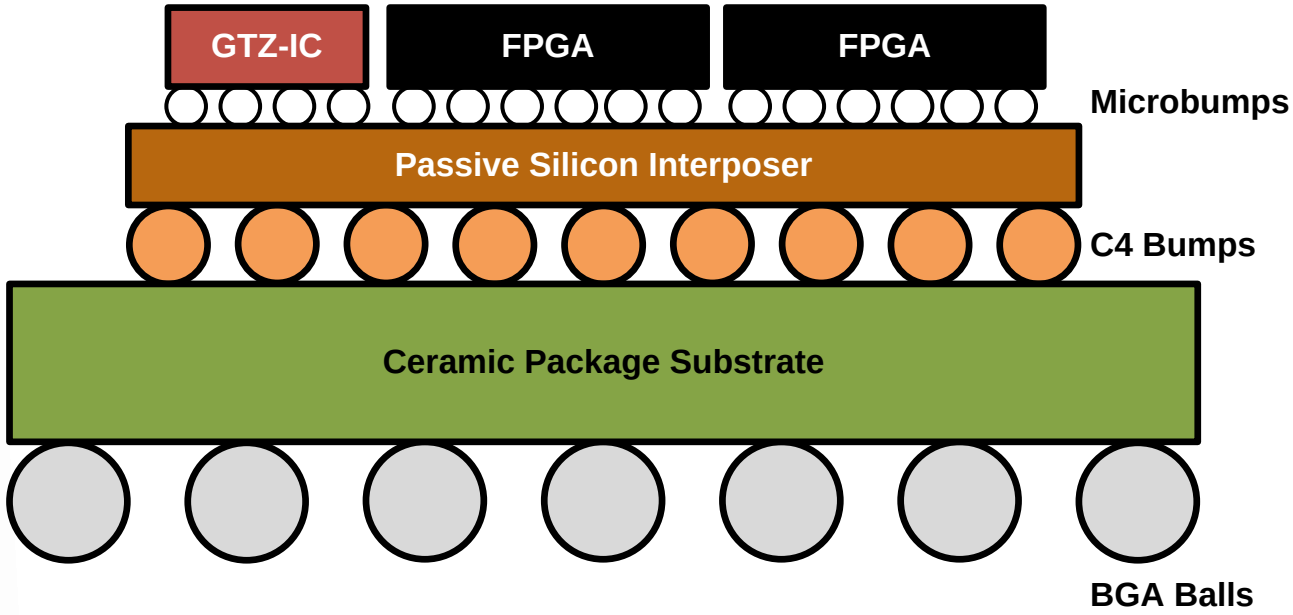
XILINX ALL PROGRAMMABLE

XC7VH580T Under the Hood

Industry's First Heterogeneous FPGA



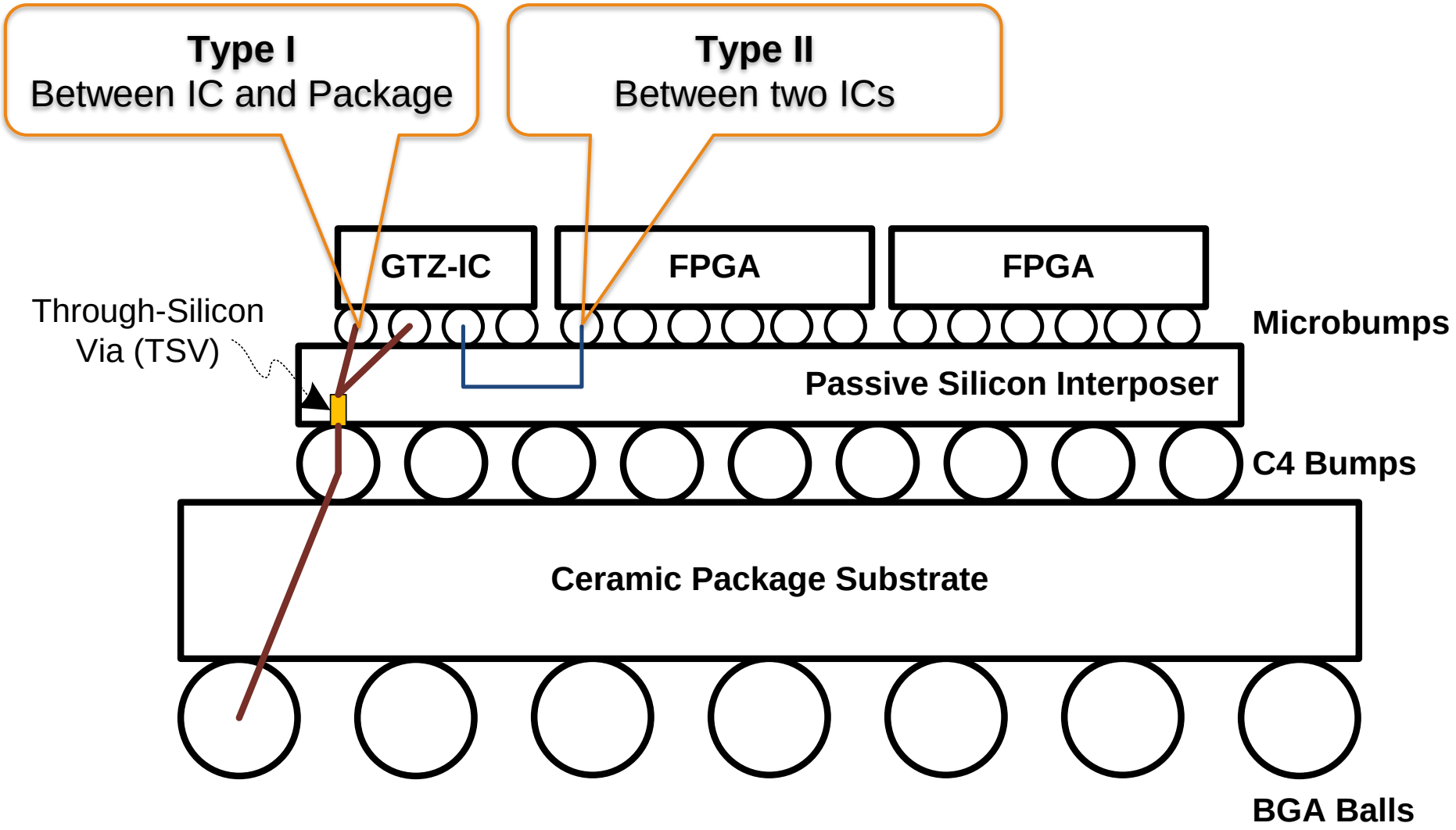




© Copyright 2012 Xilinx

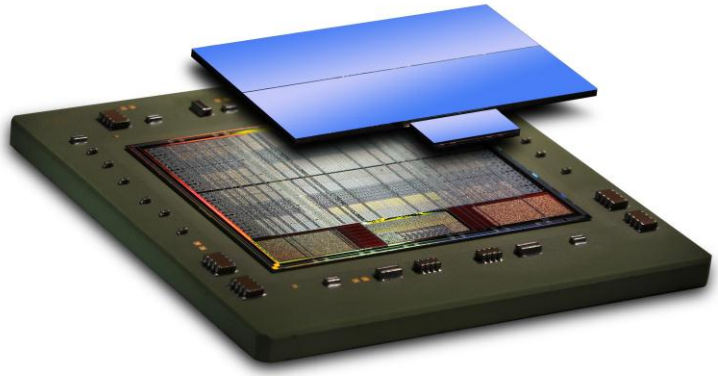
XILINX ALL PROGRAMMABLE

Two Interconnect Types



© Copyright 2012 Xilinx
XILINX ALL PROGRAMMABLE

Packaging, Assembly, and Test



Top Dice, Interposer, & Package

XILINX

28nm FPGA & Interposer

tsmc

Package Substrate

KYOCERA

µBump, Die separation
Joining, & Assembly

Amkor Technology

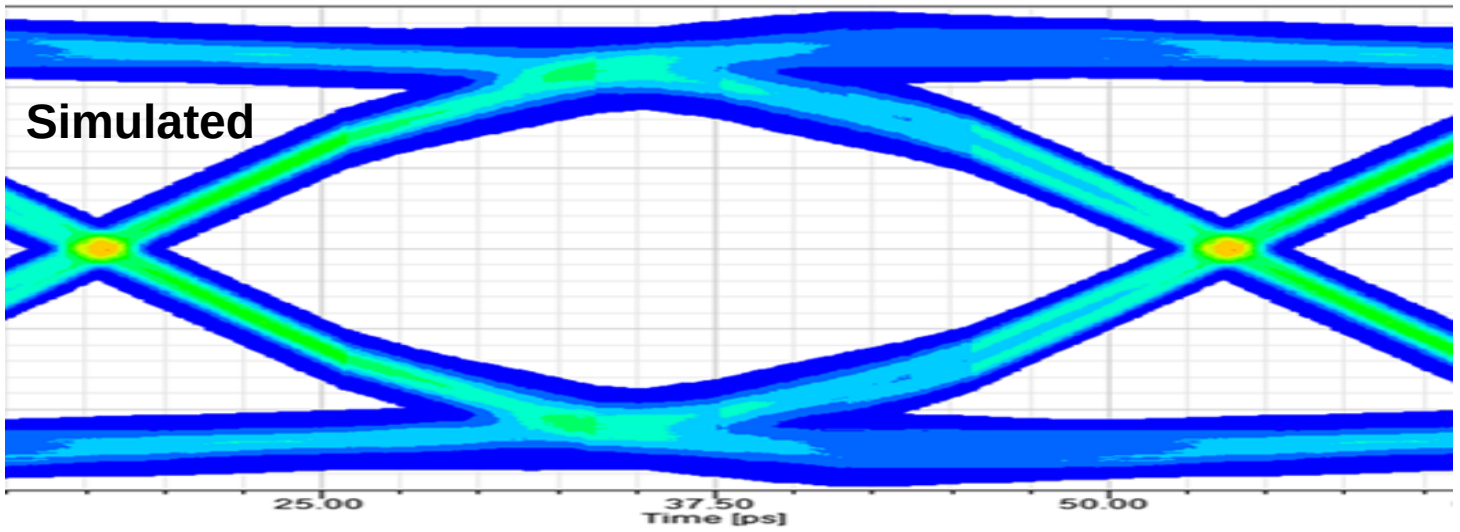
tsmc

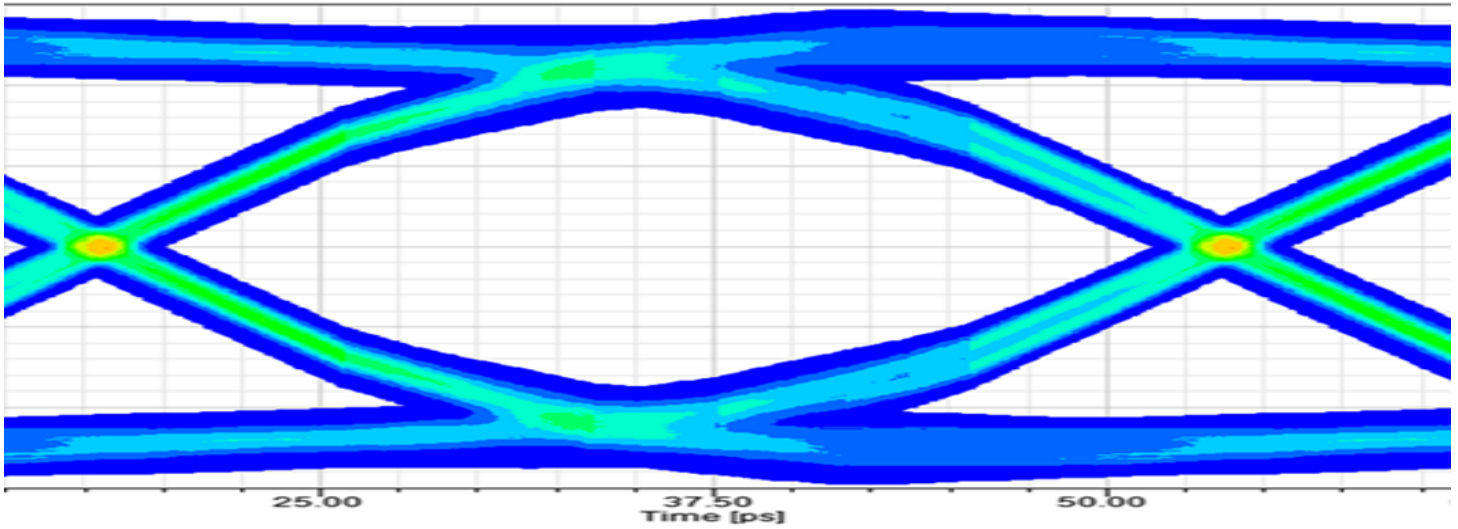
Final Test of Packaged Part

XILINX

© Copyright 2012 Xilinx

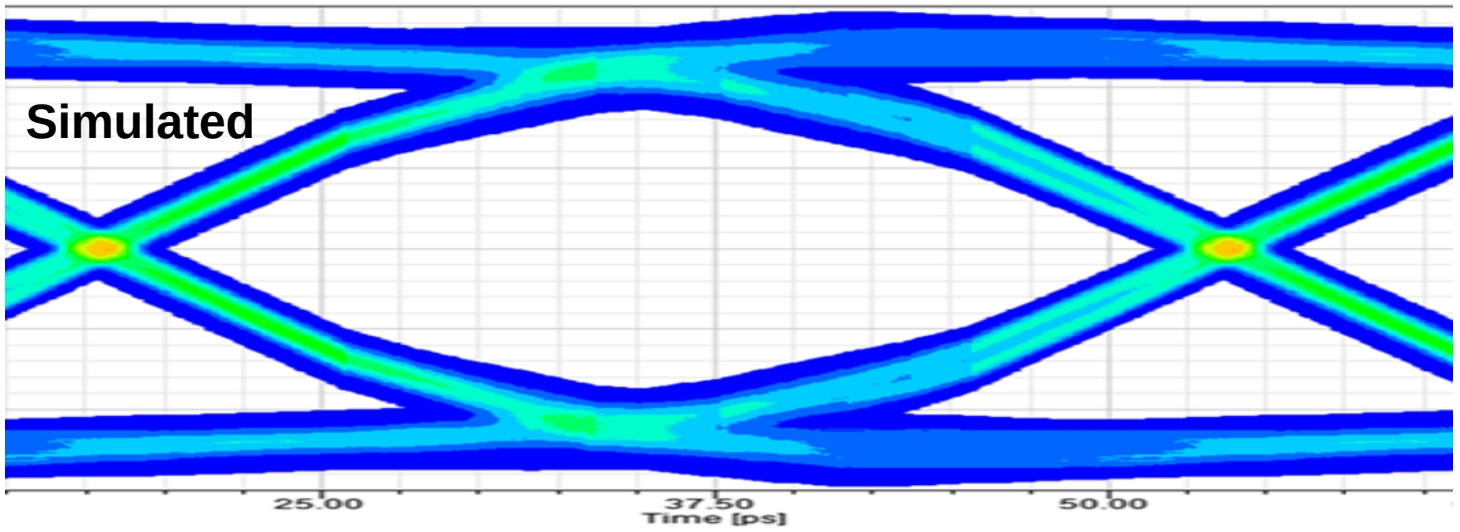
Type I Example: 28 Gb/s Serial Transmitter

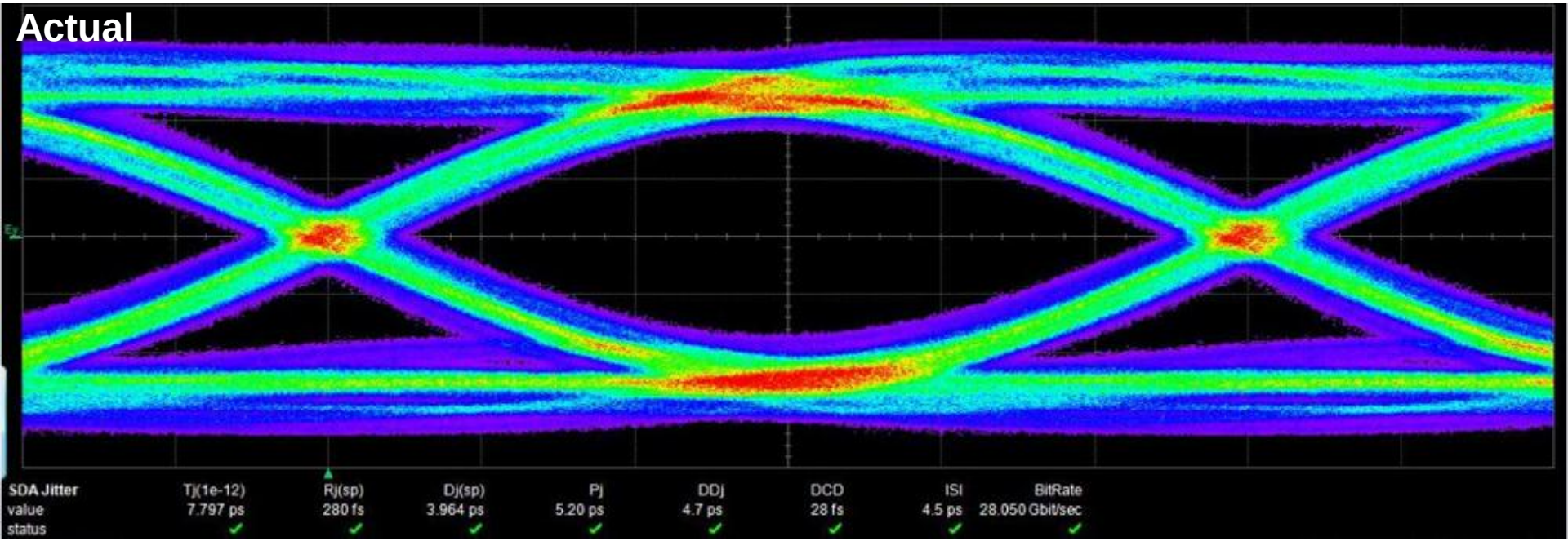




© Copyright 2012 Xilinx

Type I Example: 28 Gb/s Serial Transmitter



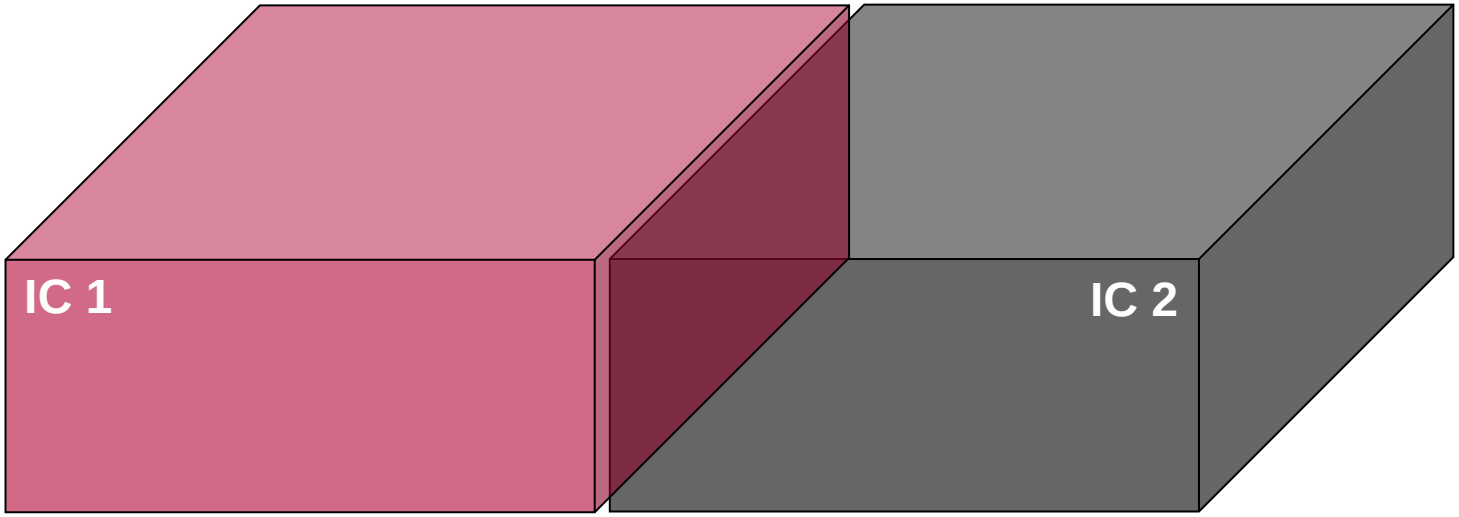


© Copyright 2012 Xilinx

XILINX ALL PROGRAMMABLE.

Type II Interconnects

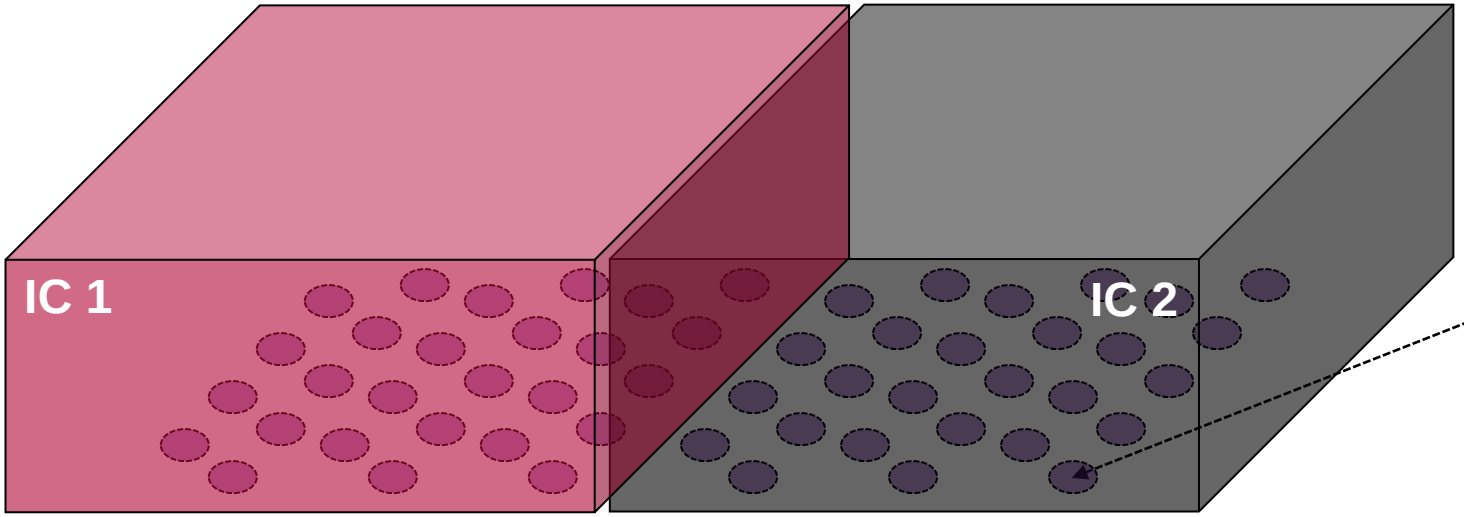
Inter-IC Interconnect Microstrip Layout with Side Shields



© Copyright 2012 Xilinx
XILINX ALL PROGRAMMABLE

Type II Interconnects

Inter-IC Interconnect Microstrip Layout with Side Shields

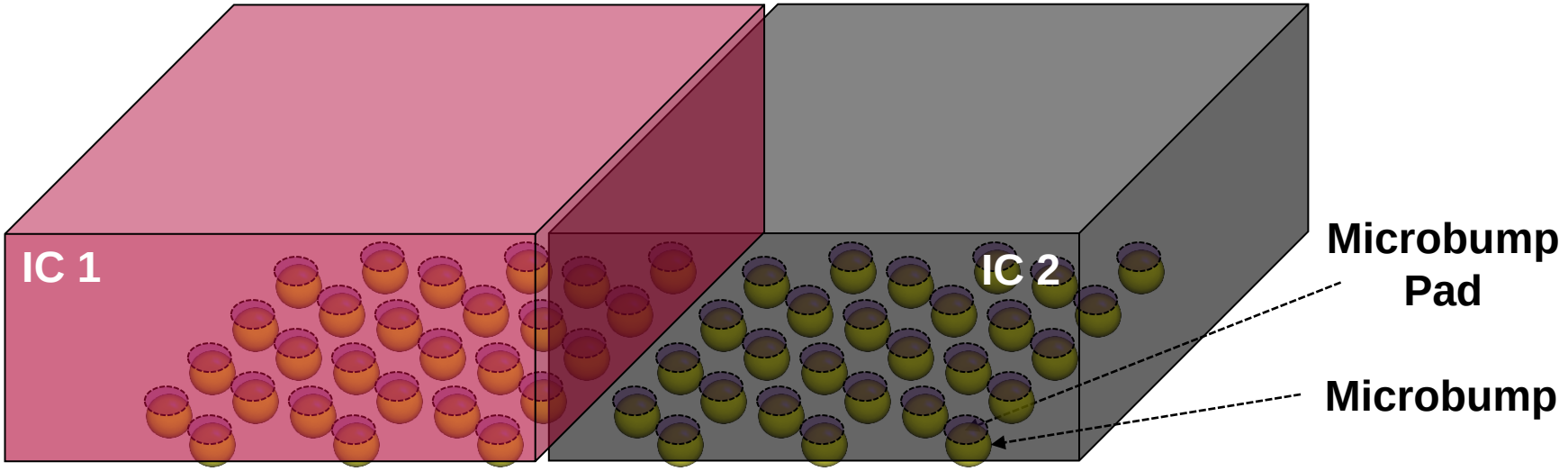


© Copyright 2012 Xilinx

XILINX ALL PROGRAMMABLE

Type II Interconnects

Inter-IC Interconnect Microstrip Layout with Side Shields

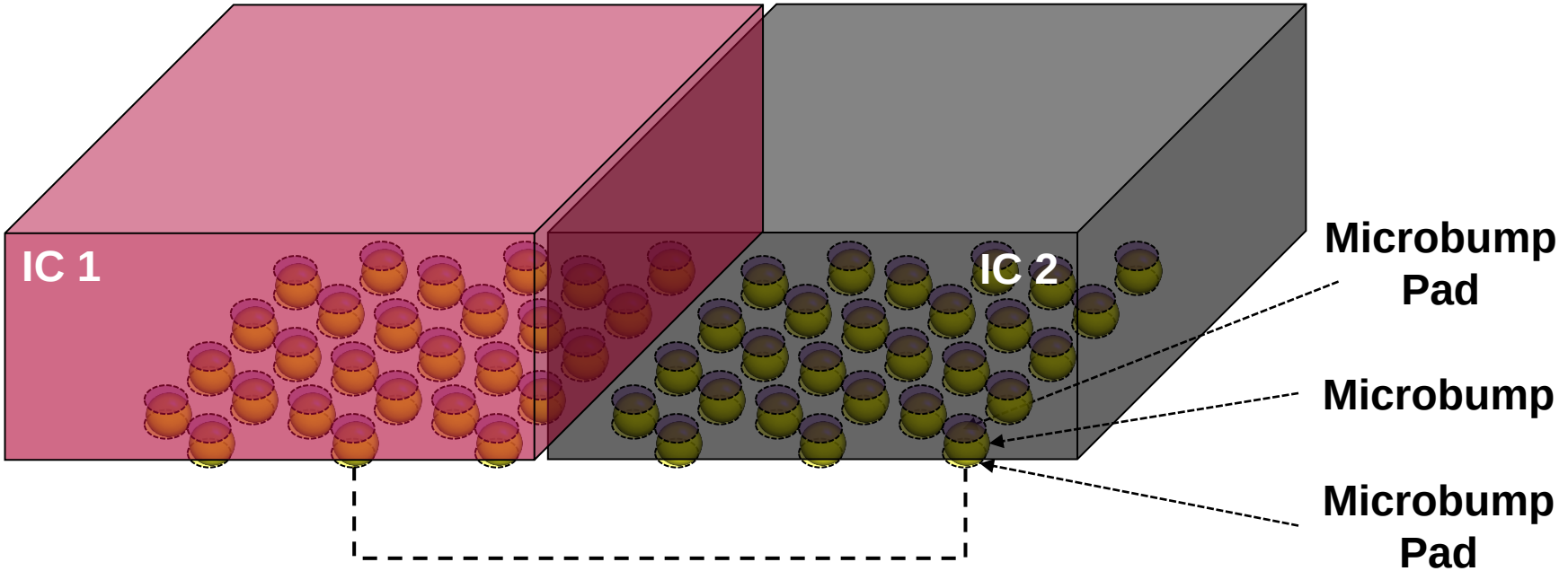


© Copyright 2012 Xilinx

XILINX ALL PROGRAMMABLE

Type II Interconnects

Inter-IC Interconnect Microstrip Layout with Side Shields

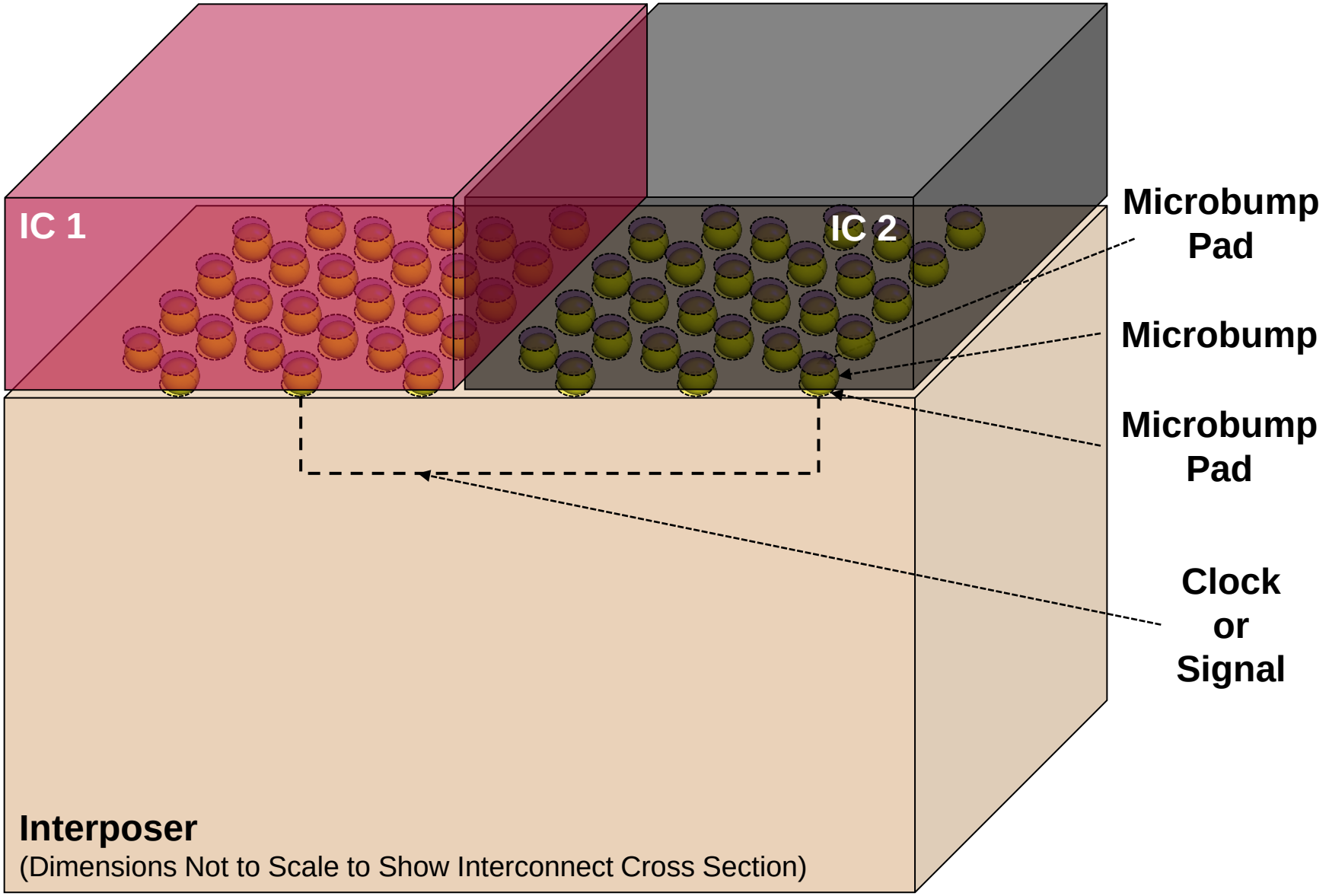


© Copyright 2012 Xilinx

XILINX ALL PROGRAMMABLE

Type II Interconnects

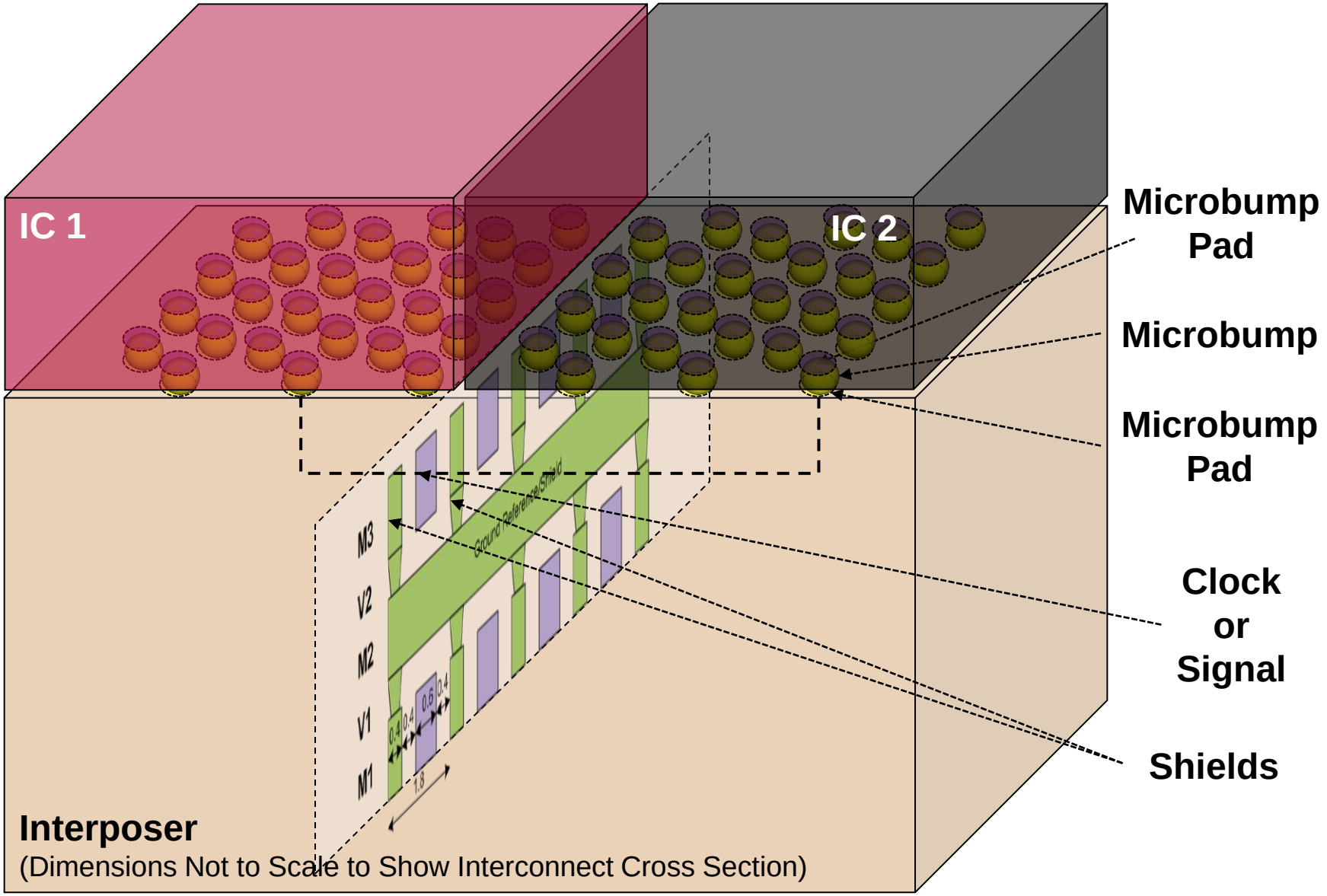
Inter-IC Interconnect Microstrip Layout with Side Shields



© Copyright 2012 Xilinx
XILINX ALL PROGRAMMABLE

Type II Interconnects

Inter-IC Interconnect Microstrip Layout with Side Shields

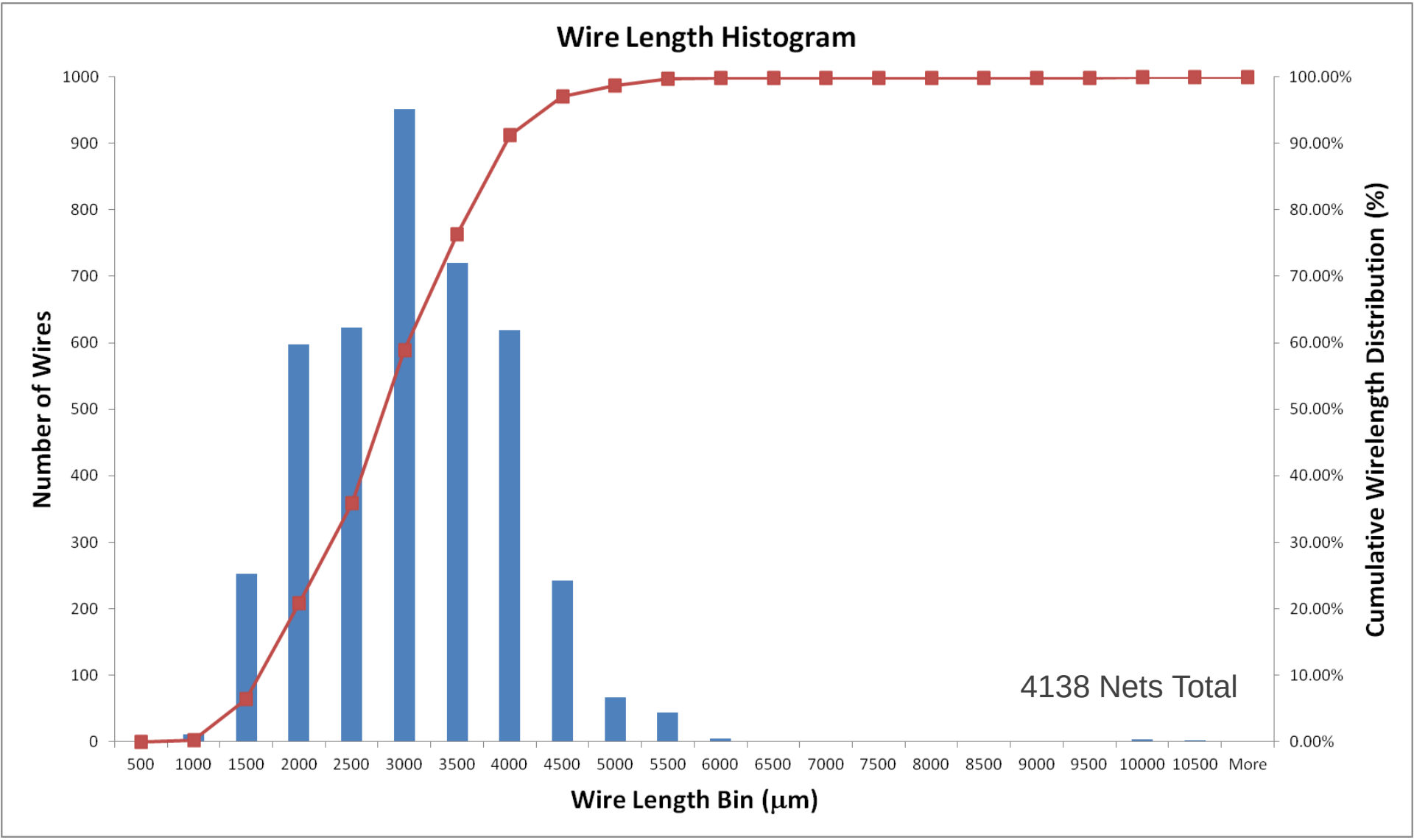


© Copyright 2012 Xilinx

XILINX ALL PROGRAMMABLE

Wire Length Distribution

Between GTZ-IC and FPGA

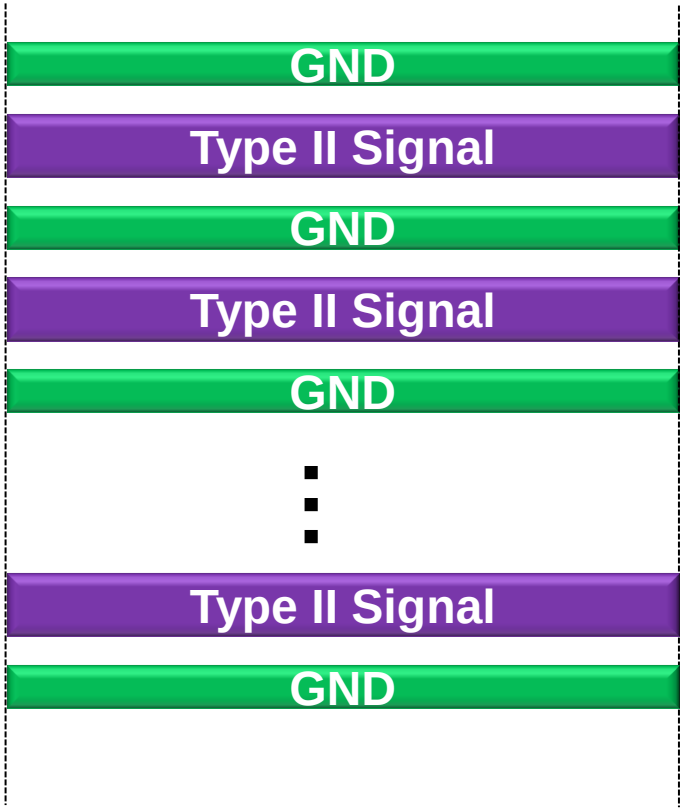


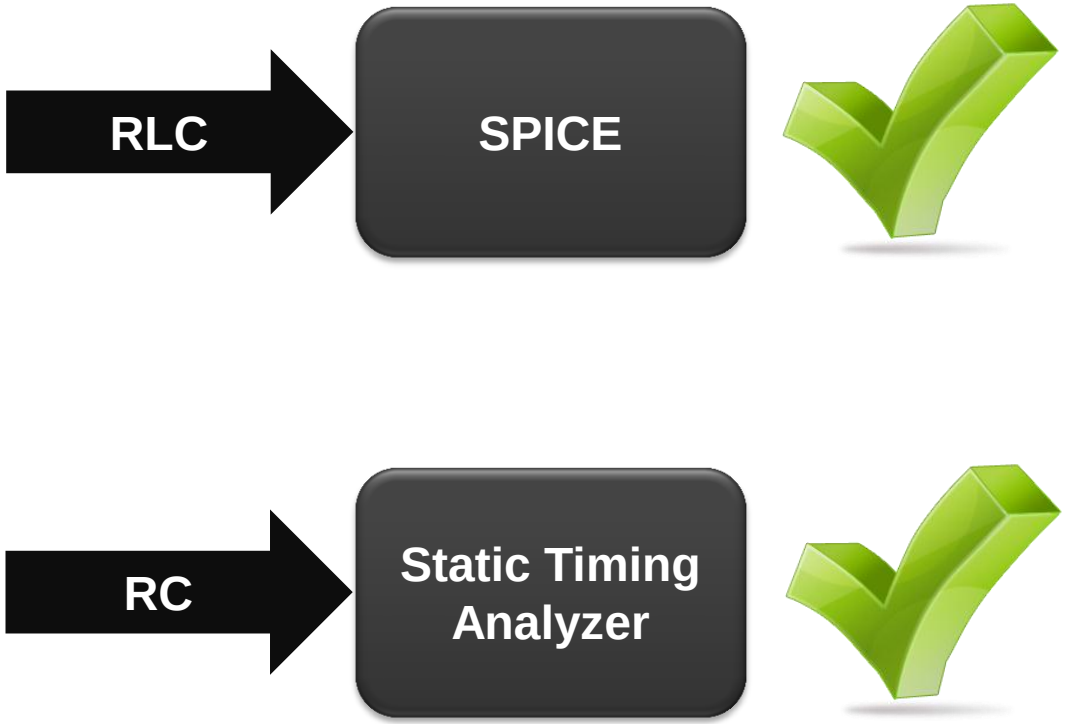
© Copyright 2012 Xilinx

XILINX ALL PROGRAMMABLE

RC Static Timing Analysis for Productivity

Calibrated RC-Based STA Against RLC-Based SPICE





© Copyright 2012 Xilinx

XILINX ALL PROGRAMMABLE

Summary

1. Presented industry's first heterogeneous 3D FPGA.
2. FPGA & GTZ-IC create three scalable products.
3. Reviewed stacked-silicon packaging & supply chain.
4. Showed Type I signaling: 28 Gb/s TX over TSVs.
5. Lacked 3D timing tools for Type II signals.
 Leveraged STA tools calibrated with RLC SPICE runs.

© Copyright 2012 Xilinx



THE FUTURE OF WIRELESS NETWORKING

Marcus Weldon

CTO Alcatel-Lucent

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

WHAT IS **REALLY** DRIVING THE (WIRELESS) MARKET ?

WHERE IS THE **REAL** VALUE ?

THE NEW **REALITY**

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

WHAT IS **REALLY** DRIVING THE (WIRELESS) MARKET ?

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE TABLET GENERATION IS IN COMMAND

67%
Would cut anything but **Mobile BB** (UK)

70%
Mobile-only Web users in emerging markets

100%
Broadband users **microblog** (China)

500M+
Users/month on **Facebook** apps platform

84%
Choose **Internet** over partner or car (Germany)



11.5
Content hours in 7hrs by 8-18 years old (USA)

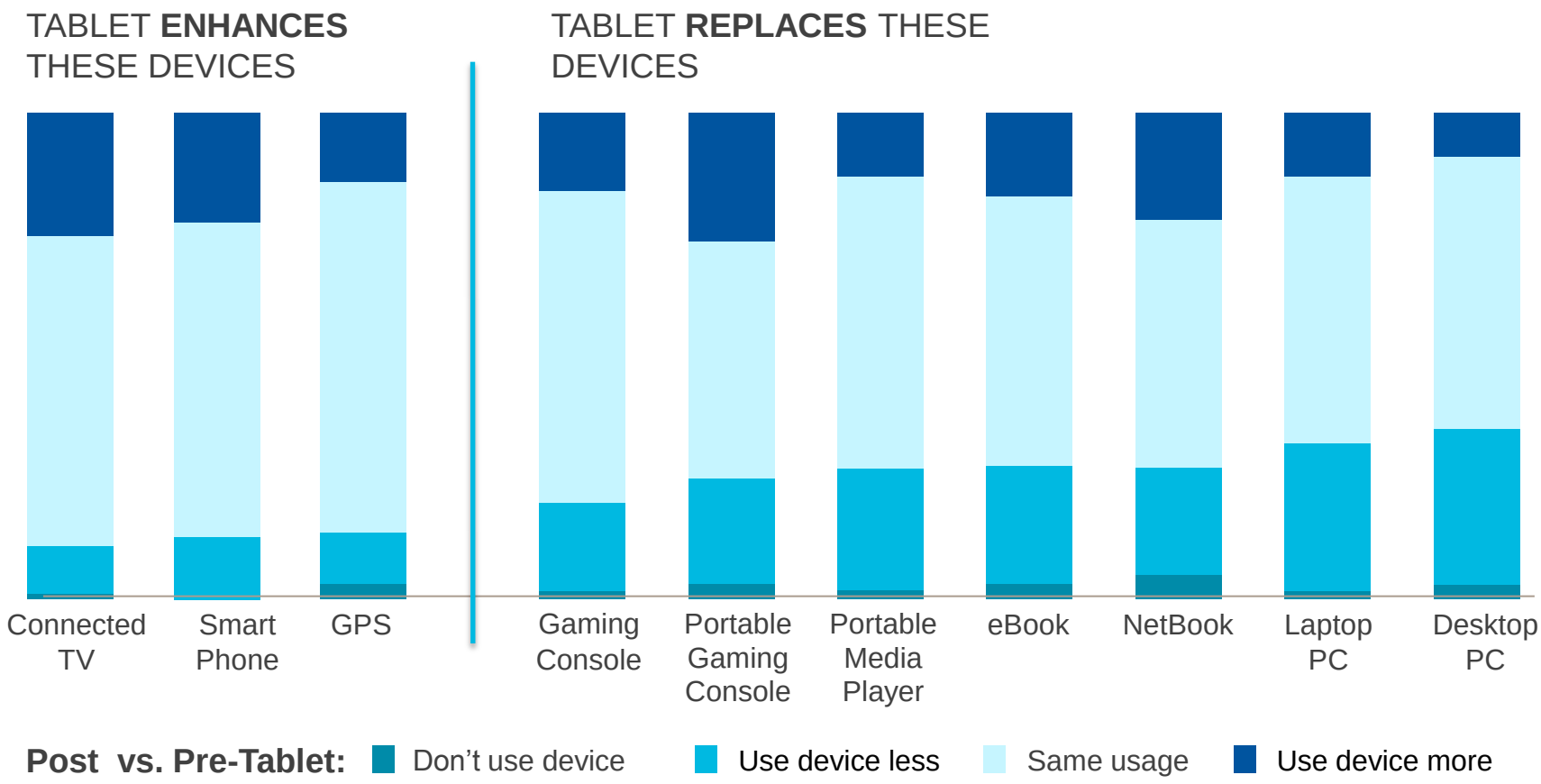
66%
Sleep with **smart phone** (USA)

100M+
Tablets sold in 2012 globally

USERS KNOW WHAT THEY WANT AND HOW IT SHOULD BE DELIVERED

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE TABLET VERSUS. . .
EVERYTHING ELSE

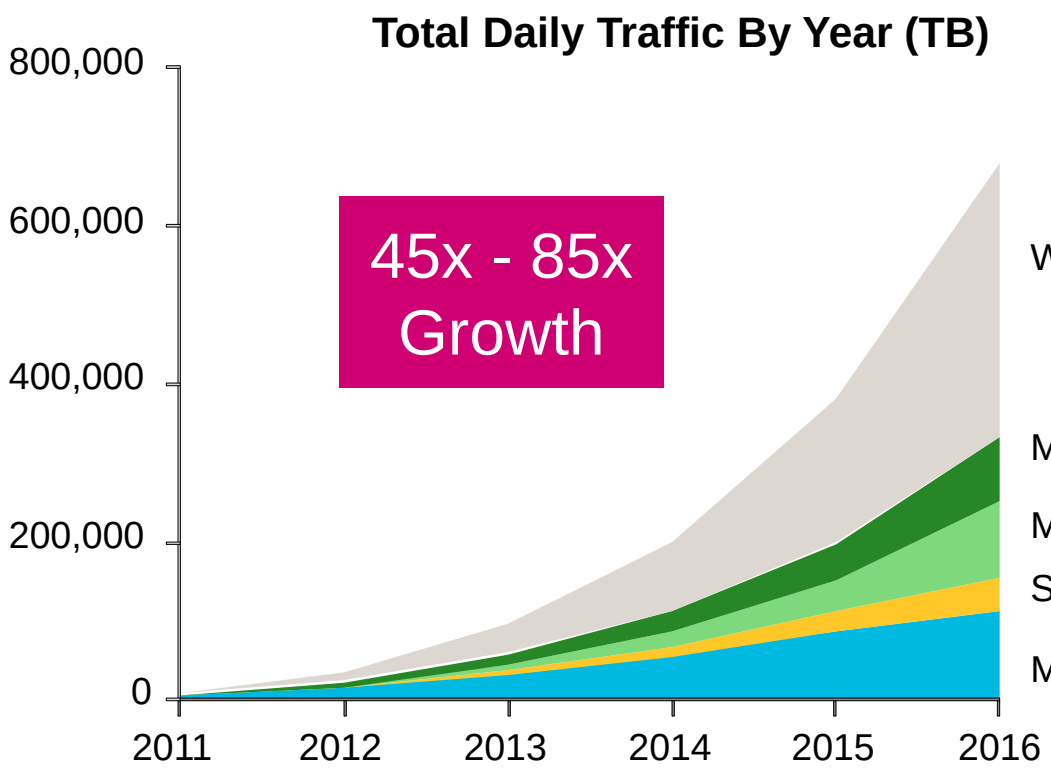


LIFE CONVERGENCE: WORK/HOME, CELL/WIFI, EVERYTHING EVERYWHERE

Source: The Nielsen Company

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE NET EFFECT:
THIS IS VERY *DEMANDING*



Bell Labs Modeling

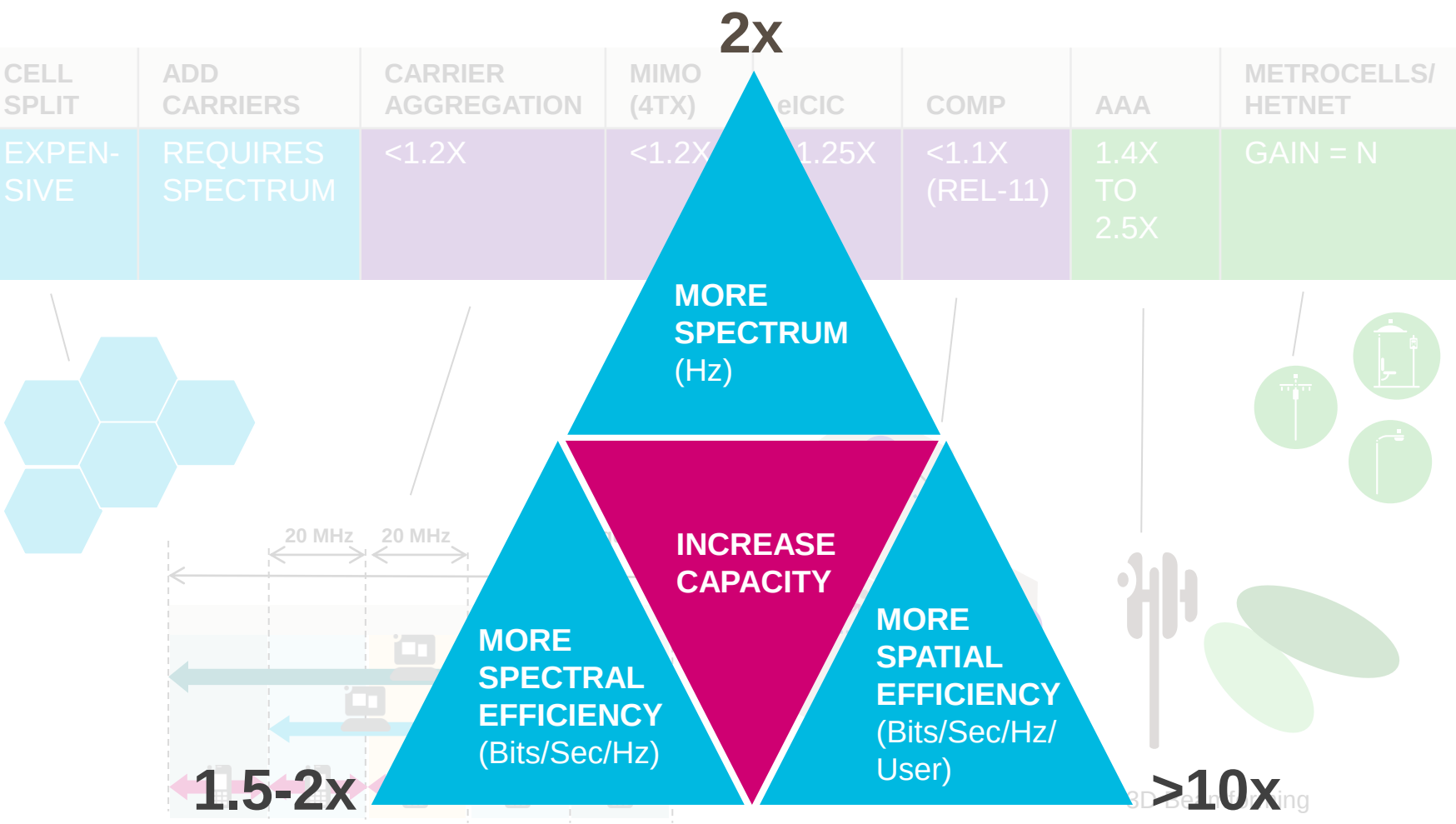
	CAGR 2011- 2016	
Metro Cell (WiFi)	24%	54%
Metro Cell (3G/LTE)	30%	
Smartload	13%	
Macro	34%	

MASSIVE GROWTH IN DEMAND REQUIRES NEW SUPPLY STRATEGY

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE SUPPLY LEVERS:
RADIO CAPACITY

CELL SPLIT	ADD CARRIERS	CARRIER AGGREGATION	MIMO (4TX)	eICIC	COMP	AAA	METROCELLS/ HETNET
EXPEN-SIVE	REQUIRES SPECTRUM	<1.2X	<1.2X	1.25X	<1.1X (REL-11)	1.4X TO 2.5X	GAIN = N



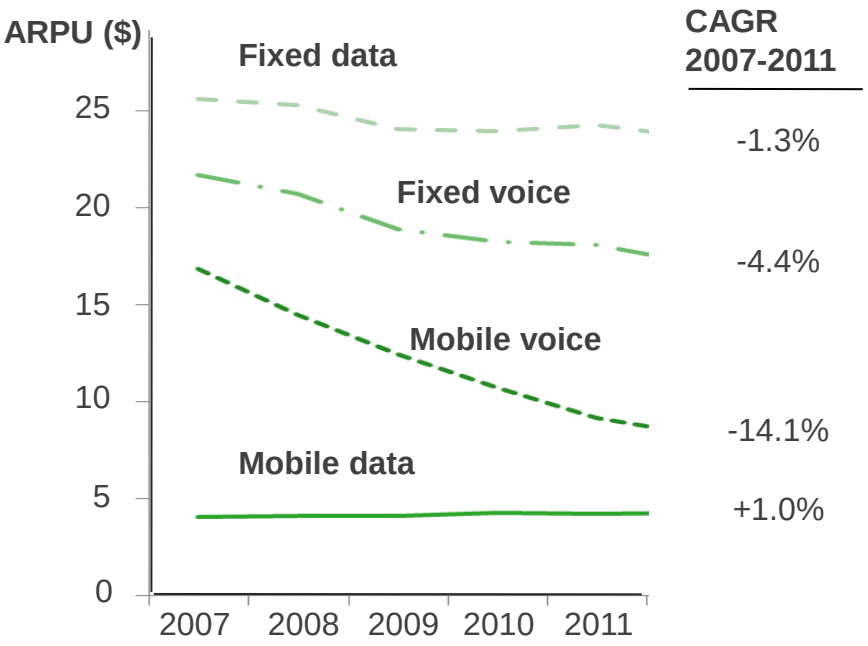
WHAT IS THE IMPACT FOR OPERATORS?

Alcatel·Lucent

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

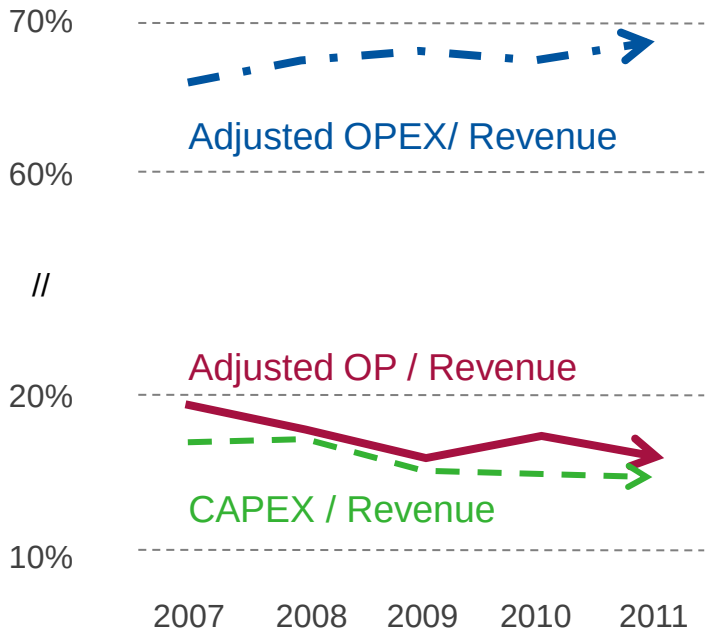
OPERATORS HAVE NOT CAPTURED THE FULL VALUE

OPERATORS CONSUMER REVENUES



+2.5% WW revenue: ARPU decline offset by more subs.

OPERATORS COSTS AND PROFITS



High pressure on CAPEX to control OP

UNLOCK THE NETWORK VALUE TO MEET THE USER DEMAND

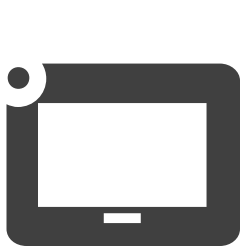
Source: Alcatel-Lucent analysis

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

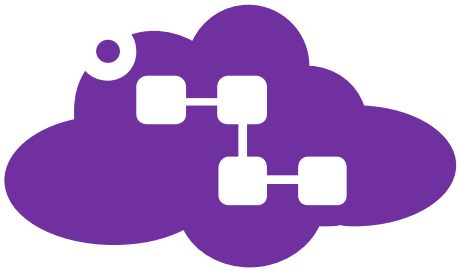
WHERE IS THE **REAL** VALUE ?

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE ESSENTIAL BRIDGE BETWEEN HAND AND CONTENT



Handheld



Cloud Network



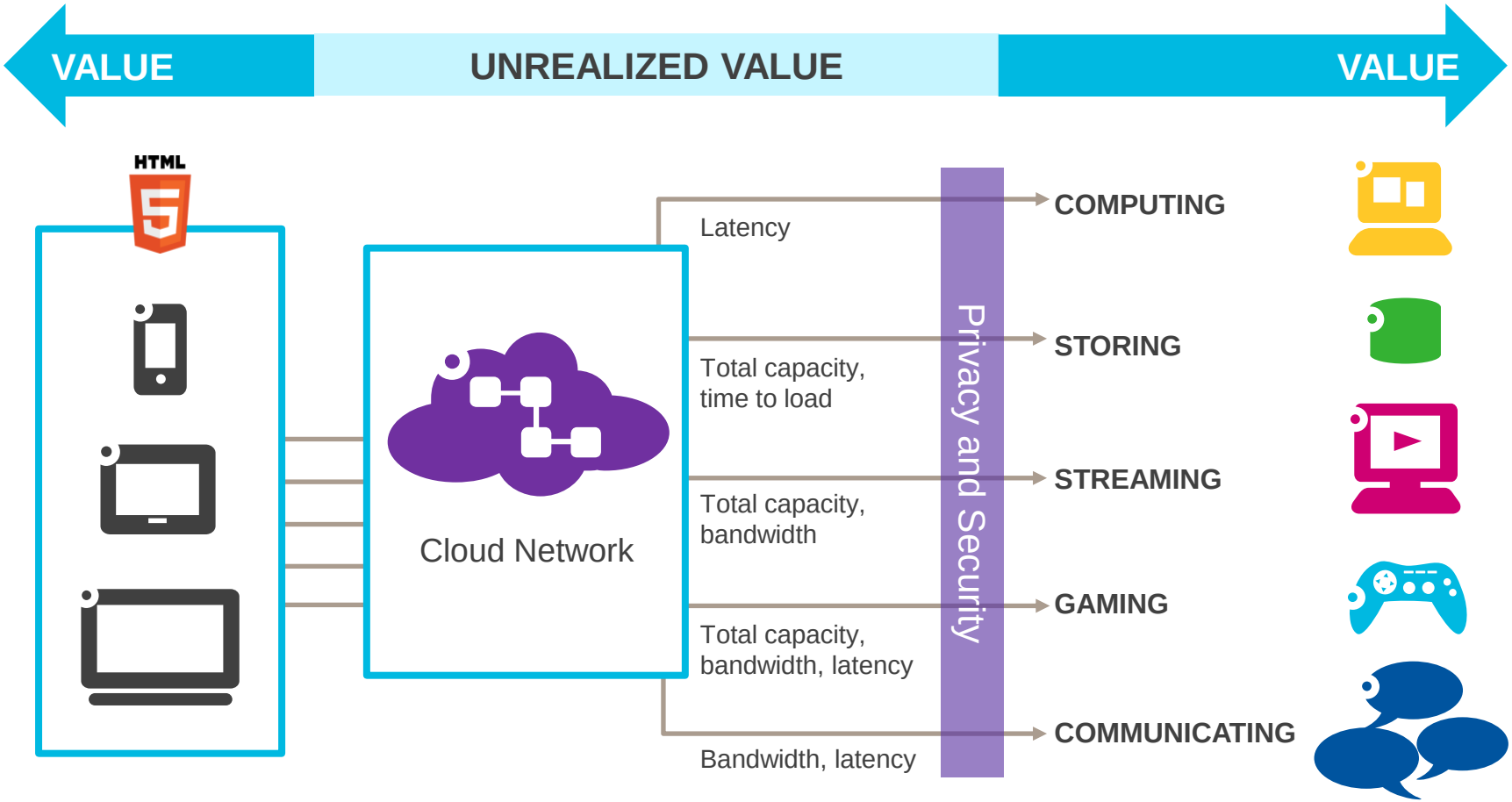
Content in cloud

BEST EXPERIENCE **BEST DELIVERY** **BEST MEDIA AND APPS**

THREE DISTINCT FUNCTIONS BUT A SINGLE UNIVERSE

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE NETWORK IS AT THE EPICENTER AND MATTERS MORE THAN EVER



THE NETWORK (AND THE OPERATOR) IS CRITICAL TO THE EXPERIENCE

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION



COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE PROFOUND SHIFT DRIVES GLOBAL GROWTH

"ANALOG" ECONOMY

BOOKS | MUSIC | VIDEO | TELEPHONY | CONSUMER GOODS | ENERGY

VERTICALIZED SECTORS

- Disconnected
- Unit of mass markets
- Subscriptions, brand loyalty
- Hardware-Defined
- Proximity-based groups
- Independent economies
- Innovation timescale = years

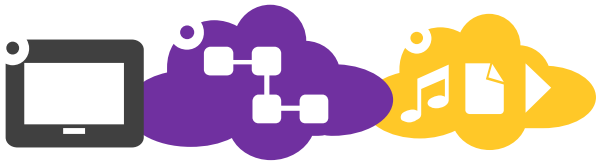
"DIGITAL" ECONOMY



PARTIAL RE-CONSTRUCTION TO DIGITAL INDUSTRIES

- Connected
- Unit of family, friends, colleagues
- Digital cannibalization
- Software-Defined
- Rise of virtual social groups
- Interdependent markets
- Innovation timescale = months

"NEXT DIGITAL" ECONOMY



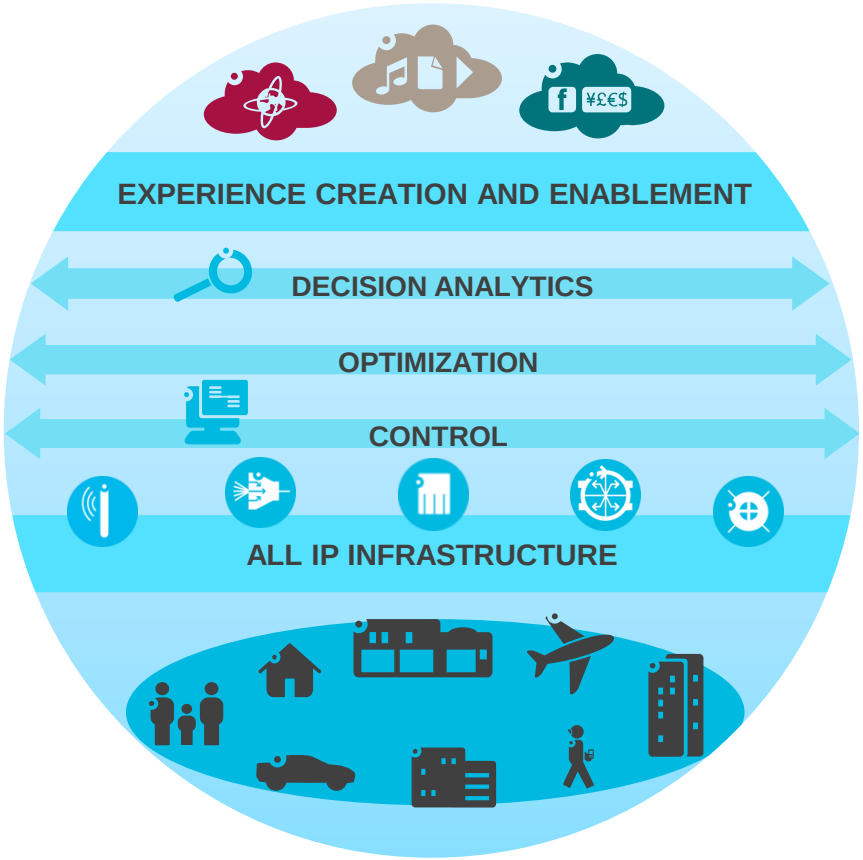
DEEP SOCIETAL CHANGE

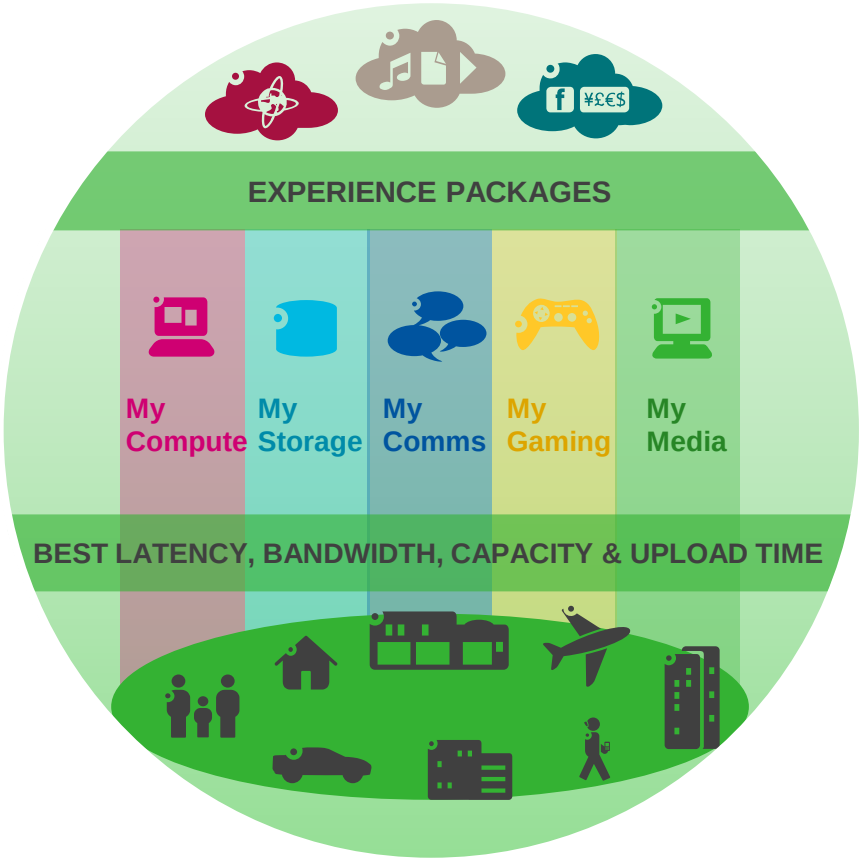
- **Hyper-connected**
- **Unit of one,** highly empowered
- **A-la-carte** user experience
- **Application-Driven**
- **Virtual** global communities dominate
- **Global** market and economy
- Innovation timescale = **days**

USER ARE MAKING THE MOVE, ARE WE READY?

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE FUTURE OF NETWORKS
A PLATFORM FOR EXPERIENCE INNOVATION

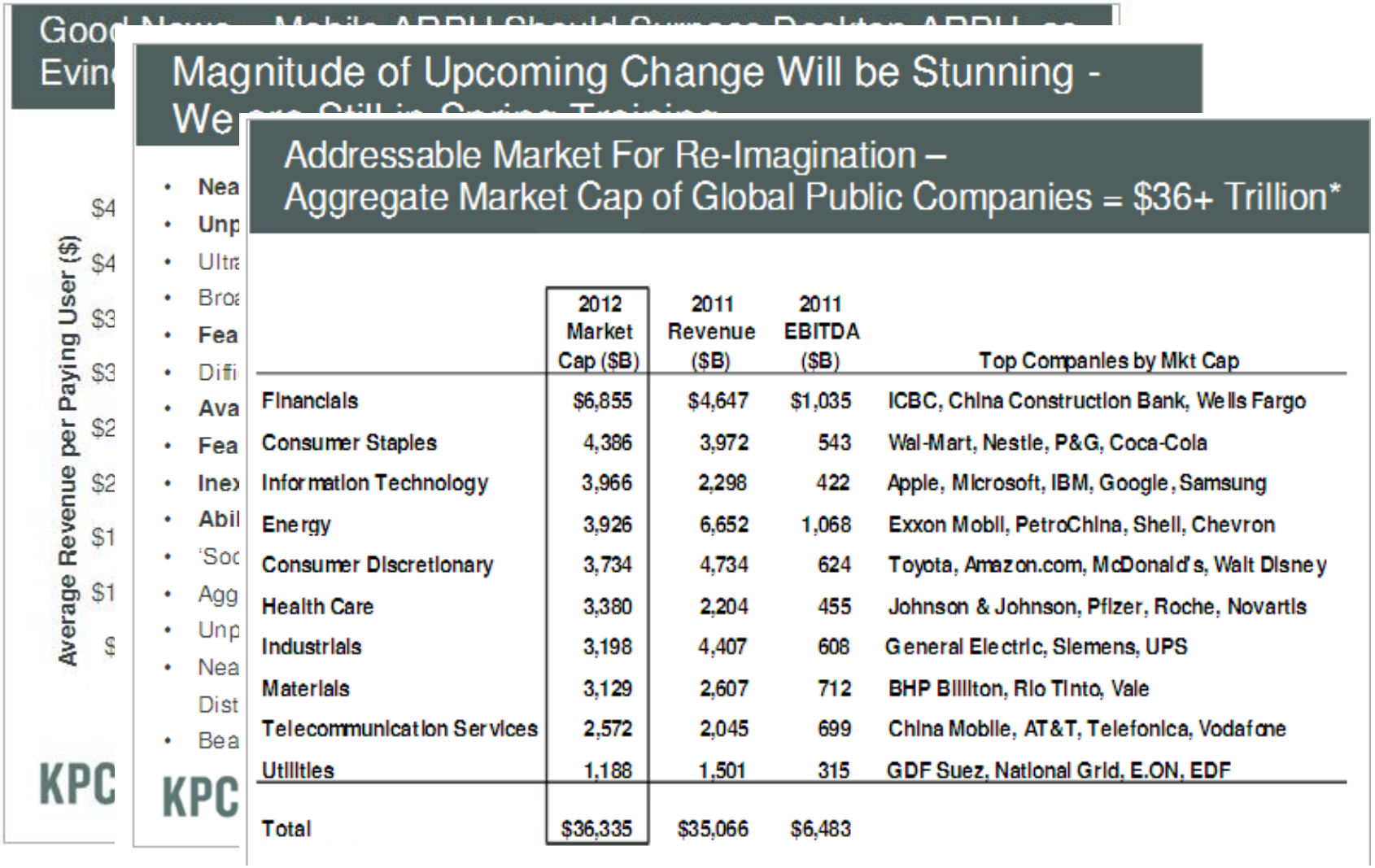




ARTIFICIAL INTELLIGENCE BUILT ON AND *IN* THE NETWORK

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE FUTURE OF NETWORKS: WHERE'S THE MONEY?



Magnitude of Upcoming Change Will be Stunning -

Addressable Market For Re-Imagination –
Aggregate Market Cap of Global Public Companies = $36+ Trillion*

	2012 Market Cap ($B)	2011 Revenue ($B)	2011 EBITDA ($B)	Top Companies by Mkt Cap
Financials	$6,855	$4,647	$1,035	ICBC, China Construction Bank, Wells Fargo
Consumer Staples	4,386	3,972	543	Wal-Mart, Nestle, P&G, Coca-Cola
Information Technology	3,966	2,298	422	Apple, Microsoft, IBM, Google, Samsung
Energy	3,926	6,652	1,068	Exxon Mobil, PetroChina, Shell, Chevron
Consumer Discretionary	3,734	4,734	624	Toyota, Amazon.com, McDonald's, Walt Disney
Health Care	3,380	2,204	455	Johnson & Johnson, Pfizer, Roche, Novartis
Industrials	3,198	4,407	608	General Electric, Siemens, UPS
Materials	3,129	2,607	712	BHP Billiton, Rio Tinto, Vale
Telecommunication Services	2,572	2,045	699	China Mobile, AT&T, Telefonica, Vodafone
Utilities	1,188	1,501	315	GDF Suez, National Grid, E.ON, EDF
Total	$36,335	$35,066	$6,483	

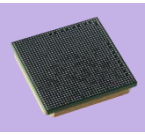
Source: Mary Meeker, KPCB, Internet Trends 2012

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE CONSTANT DRIVE IN NETWORKING
MORE EFFICIENCY, LESS SPACE, LESS ENERGY

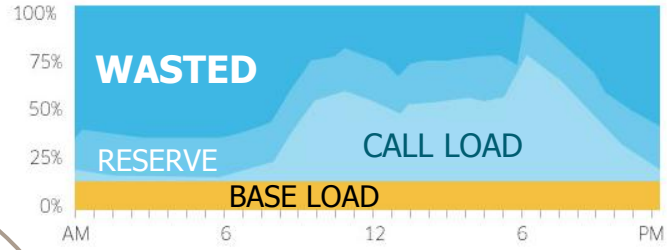
Compared to 100G systems

4X SPEED



4X DENSITY | **>2.6X** CAPACITY | **-33%** POWER

400G PIC



IMS core utilization

VIRTUALIZATION (GPP)

~ 40% less CAPACITY needed

IP CORE LEGACY





POWER 4x less
WEIGHT 10x less
SPACE 7x less

400G NP

Over past 7 years

DSL

POWER/Mbps -54%
DENSITY 3 X

VECTORING ASIC

MIGRATION: COPPER TO FIBER

GPON

POWER/Gbps -58%
DENSITY 8 X

POWER/Gbps -75%
DENSITY 10 X

LEGACY RAN



ANYG SOC

POWER SAVING -42%
TCO SAVING -51%

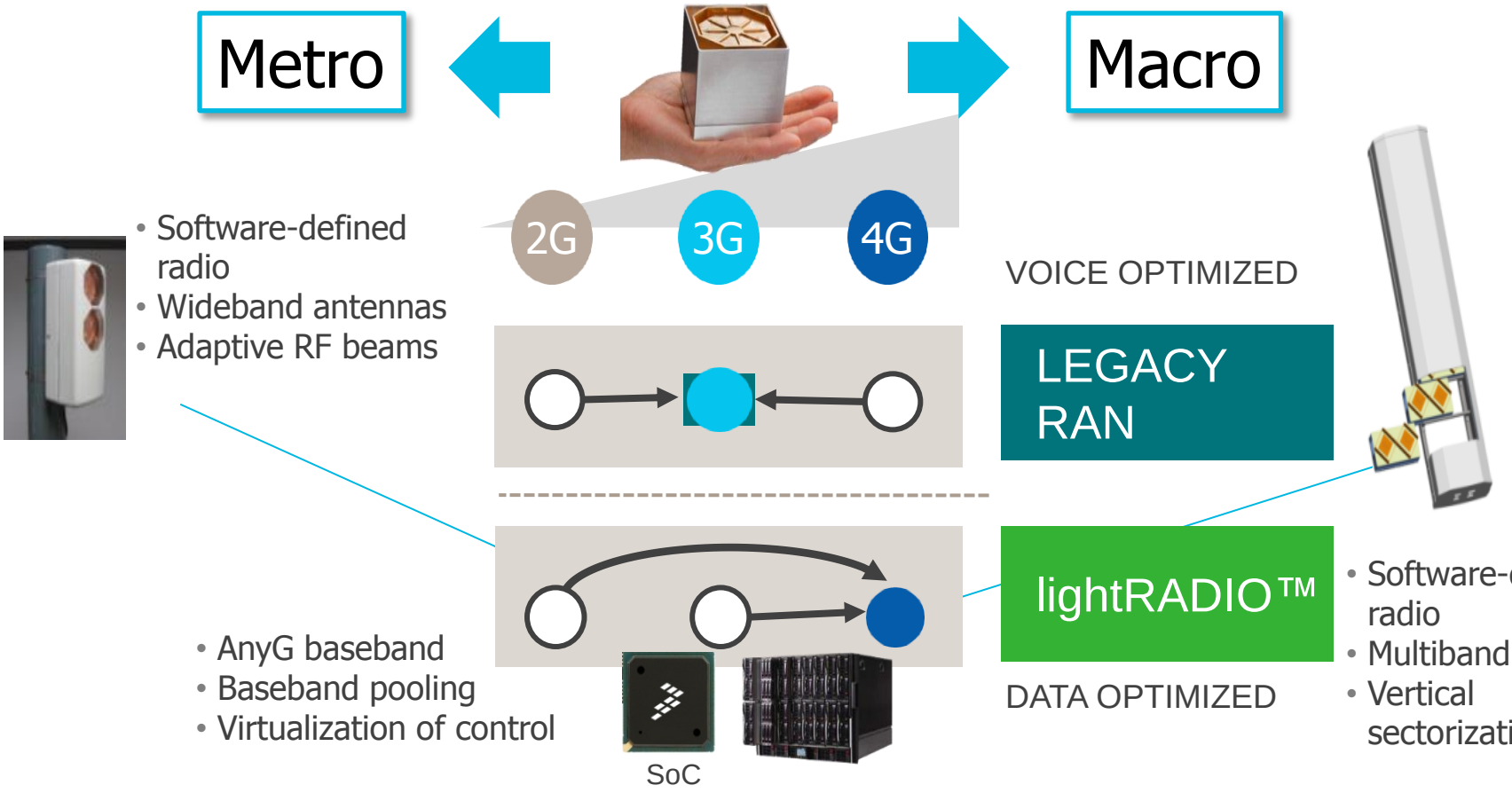


SYSTEMATICALLY PUSHING TECHNOLOGY LIMITS

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE FUTURE OF WIRELESS IS BIG & SMALL

LTE MACRO ‘OVERLAY’ AND ‘UNDERLAY’



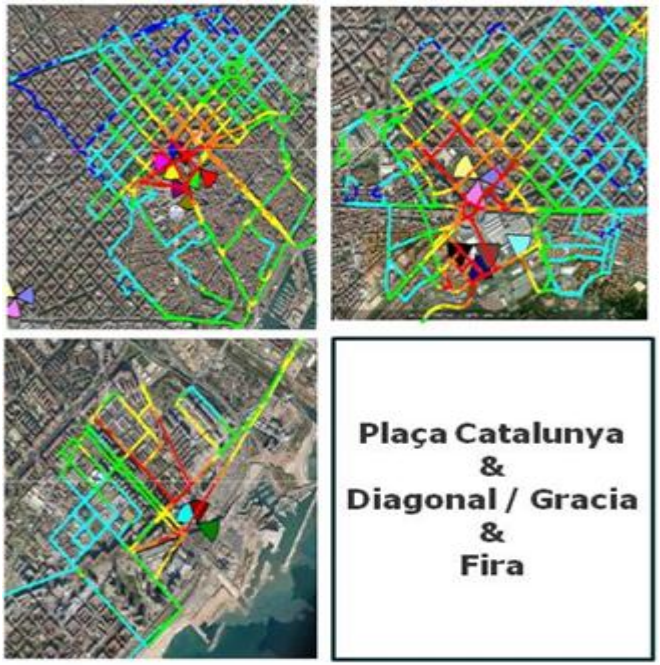
NEW 4G PLATFORMS DRIVE LOWER TCO FOR MACRO AND METRO

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE FUTURE OF WIRELESS IS BIG & SMALL
LTE METRO ‘UNDERLAY’ EXAMPLE

Delivering high capacity (100 Mbps down, 40 Mbps up) across central Barcelona

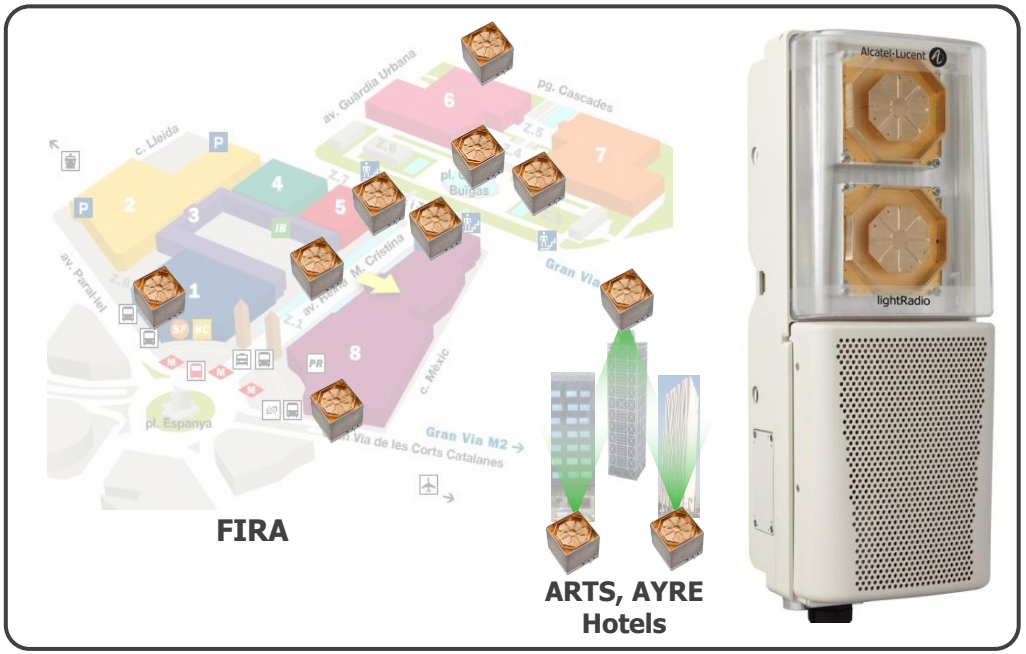
MACRO (51 Sectors)



Plaça Catalunya
&
Diagonal / Gracia
&
Fira

+

METRO (11 Sectors)



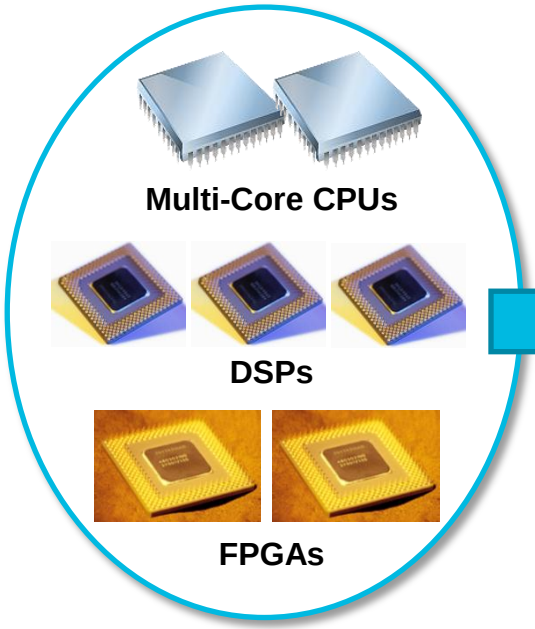
400% CAPACITY INCREASE, 40% TCO SAVINGS, 35% LESS POWER

Alcatel·Lucent

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

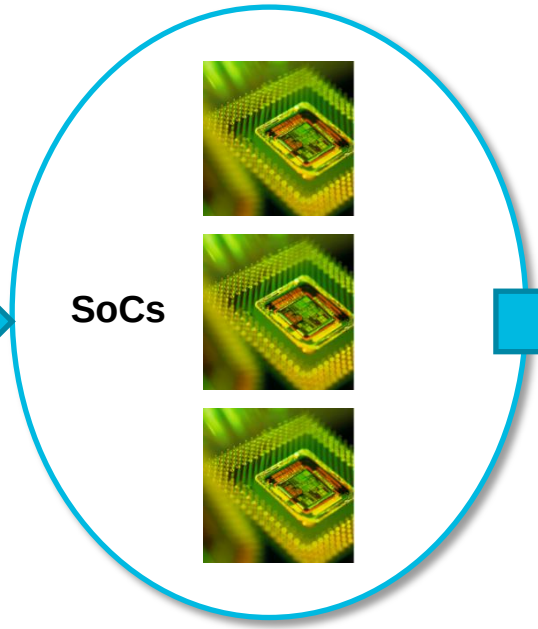
RADIO BASEBAND PROCESSING EVOLUTION
MORE EFFICIENCY, LESS SPACE, LESS ENERGY

2010 Design
Discrete L1, L2, L3 Processors



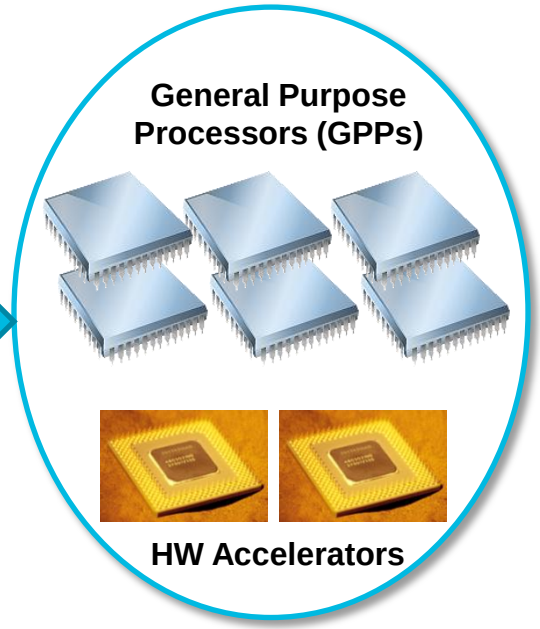
- FPGAs : L1 (PHY, Turbo Decoders,..)
- DSPs: L1 (Channel Estimation,.), and L2 (RLC/MAC, Scheduler,..)
- CPUs: L2, L3 (Transport, Security,..)

2012 Design
System On a Chip (SoC) Integrated L1, L2, L3



- HW Accelerators: L1
- Multi-Core DSPs: L1, L2
- Multi-Core CPUs: L2, L3

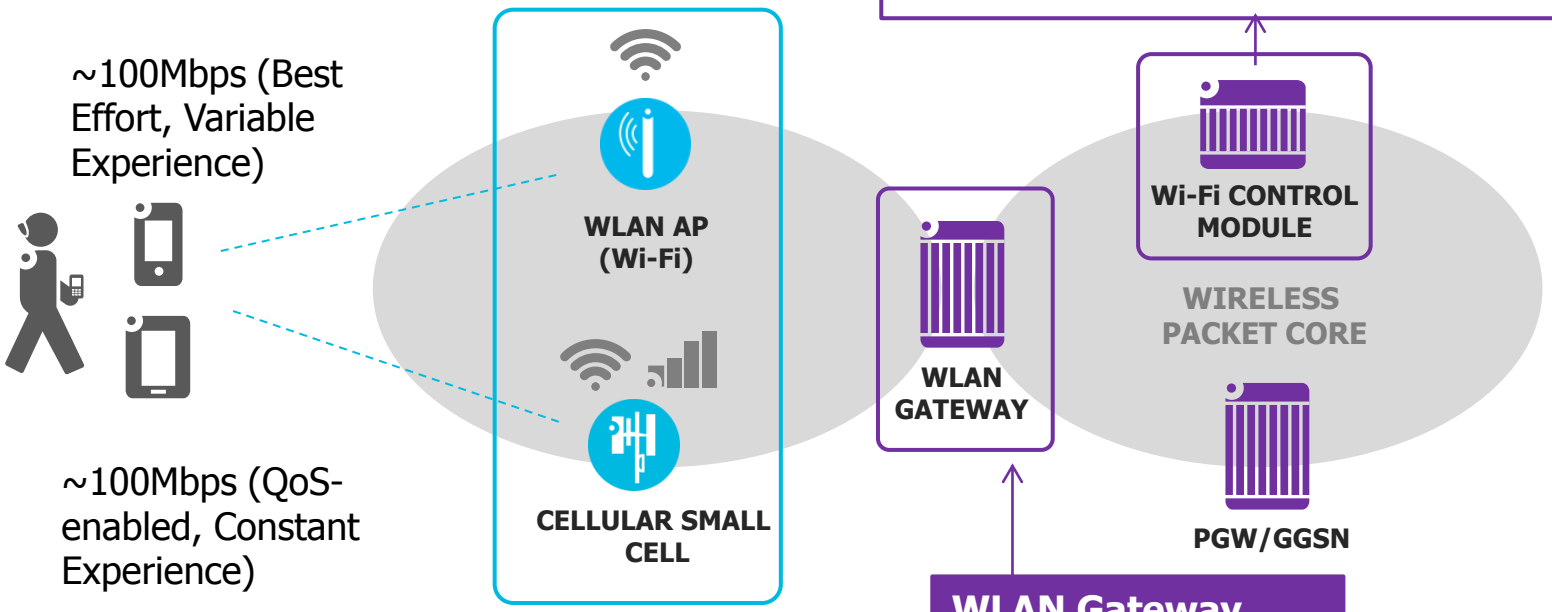
2017 Design
GPP-Centric but HW-Accelerated



- GPP: L3, L2 & some L1 Processing
- Integrated or discrete HW Accelerators: L1

SYSTEMATICALLY PUSHING TECHNOLOGY LIMITS

THE FUTURE OF WIRELESS
WHAT ABOUT WIFI?



Wi-Fi Control Module

Automatic network discovery & selection (**ANDSF**) function, enables users to be automatically connected to the "best" network

Integrated Wifi+LTE Metrocells provide 'Anyspectrum' wireless access solution, with 'Anybackhaul' option

WLAN Gateway

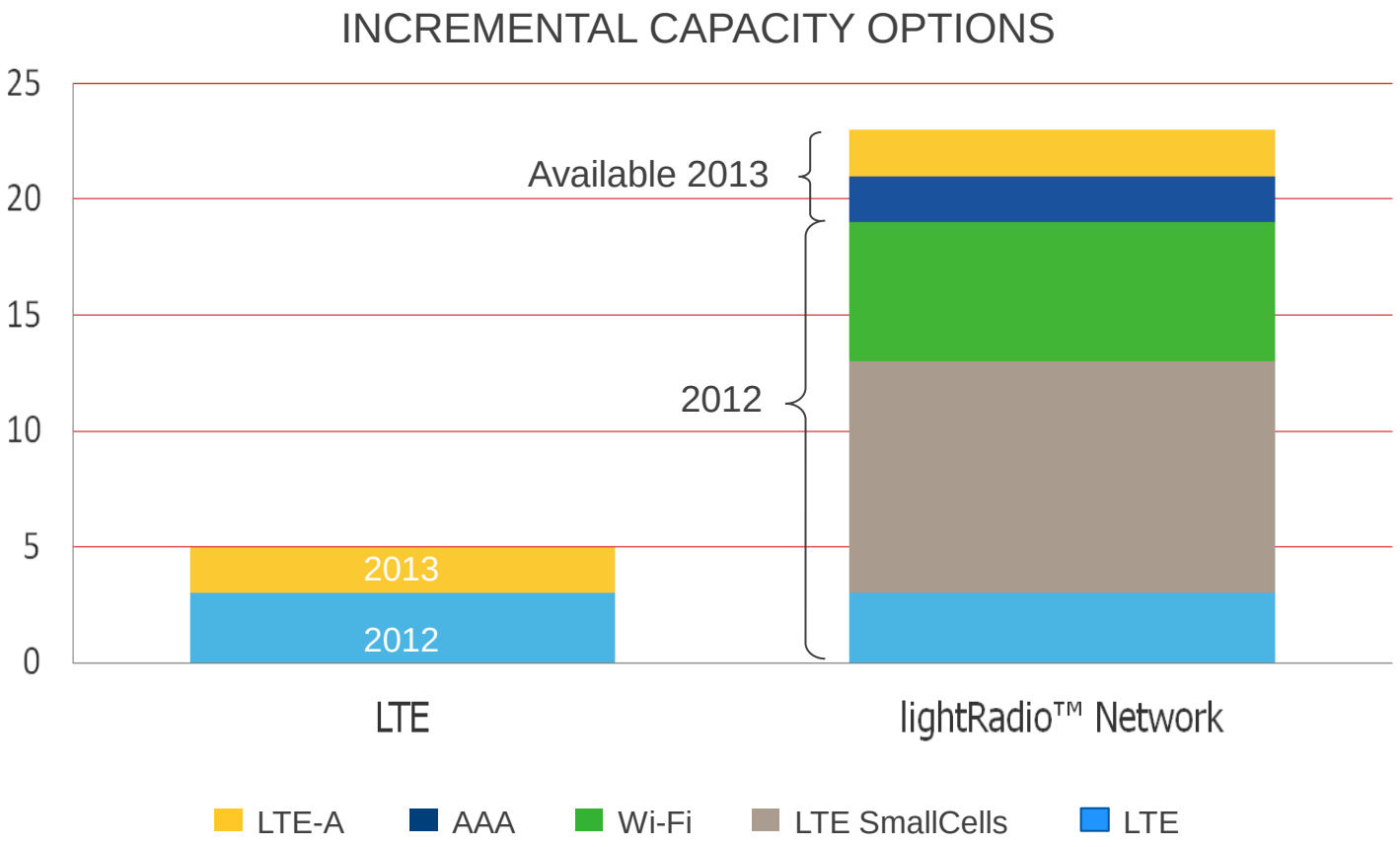
Secure Wi-Fi gateway functionality for encrypted access

SEAMLESS ROAMING BETWEEN CELLULAR AND WIFI NETWORKS BASED ON BEST NETWORK FOR APP

Alcatel·Lucent

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

THE FUTURE OF WIRELESS
SMALL CELLS CLOSE THE “DEMAND GAP”



THE FUTURE IS TECHNICALLY REALIZABLE

Alcatel·Lucent

COPYRIGHT © 2011 ALCATEL-LUCENT. ALL RIGHTS RESERVED.

IN SUMMARY

- **WHAT IS DRIVING THE MARKET?:** The Tablet Generation
- **WHERE IS THE REAL VALUE?:** Device + Cloud + *Network*
- **THE NEW REALITY:** The Network Platform (using SoCs, NPs, GPPs)



➔**THE NEXT DIGITAL ECONOMY ENABLED**

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

AT THE SPEED OF IDEAS™

COPYRIGHT © 2012 ALCATEL-LUCENT. ALL RIGHTS RESERVED.
ALCATEL-LUCENT — CONFIDENTIAL — SOLELY FOR AUTHORIZED PERSONS HAVING A NEED TO KNOW — PROPRIETARY — USE PURSUANT TO COMPANY INSTRUCTION

www.alcatel-lucent.com

Floating Point Processing using FPGAs

Michael Parker

Altera Corp

HotChips Conference

August 29, 2012

© 2011 Altera Corporation—**Confidential**

Agenda

- Stratix V FPGA architecture for Floating Point
- New Approach: “Fused Data Path”
- Throughput, GFLOPs, GFLOPs/W
 - FFT
 - Cholesky Decomposition
 - QR Decomposition
- Computational Accuracy
- Third Party Benchmarking

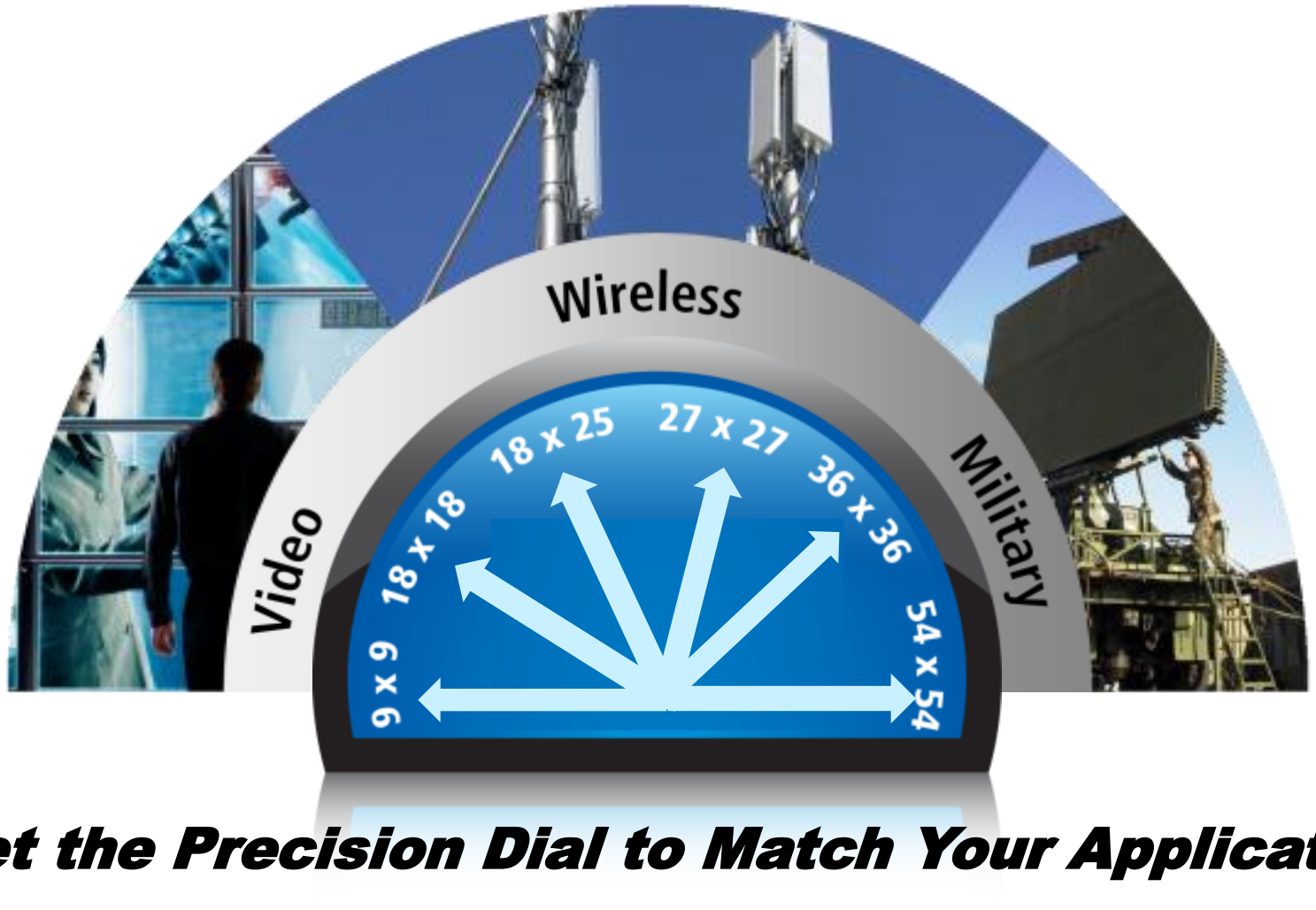
© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Stratix V architecture enhancements for floating point

© 2011 Altera Corporation—**Confidential**

Altera's Variable-Precision DSP Block



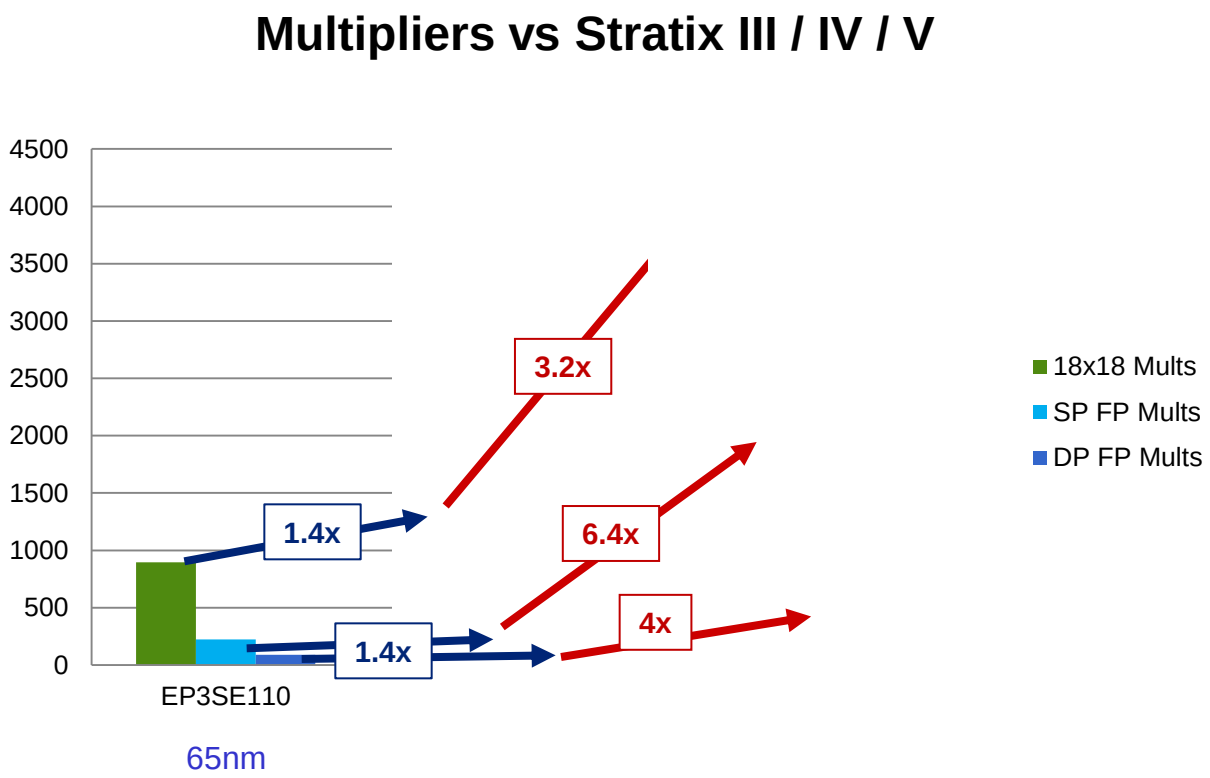
Set the Precision Dial to Match Your Application

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Why Floating Point at 28nm ?

- Floating point density determined by hard multiplier density
- Multipliers must efficiently support floating point mantissa sizes

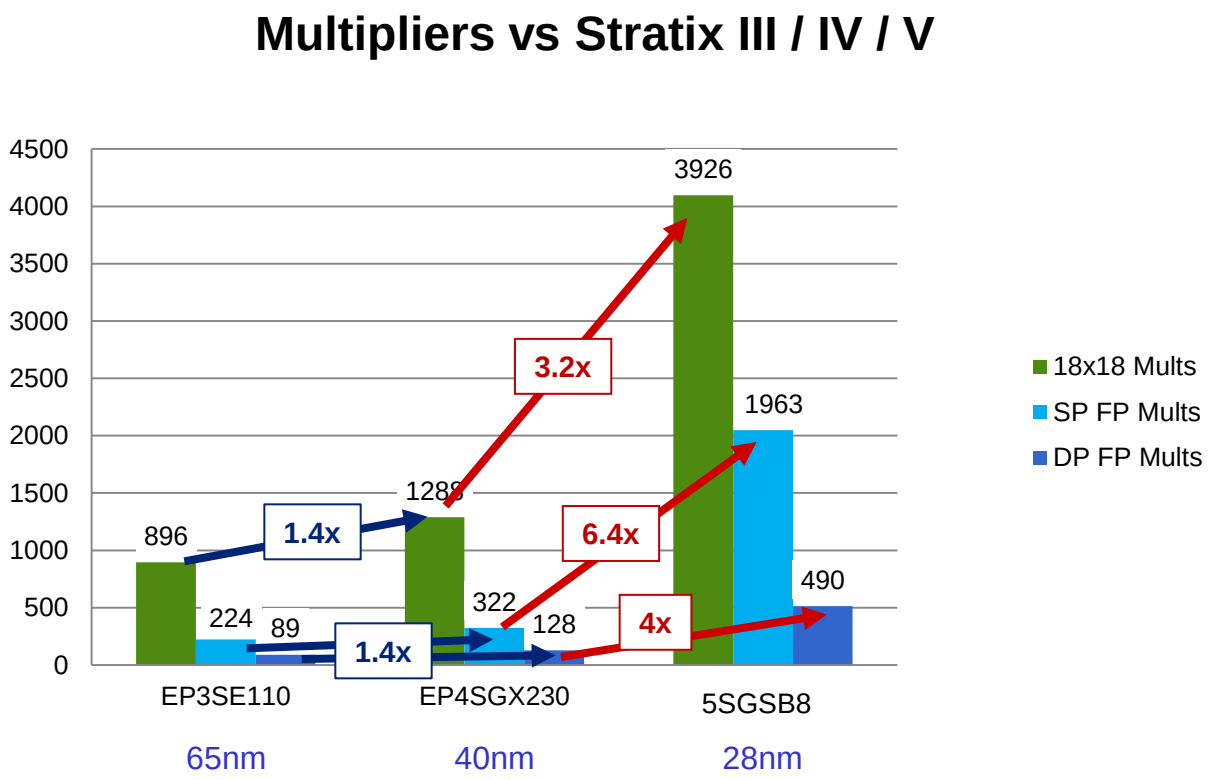


© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Floating Point Multiplier Capabilities

- Floating point density determined by hard multiplier density
- Multipliers must efficiently support floating point mantissa sizes



© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

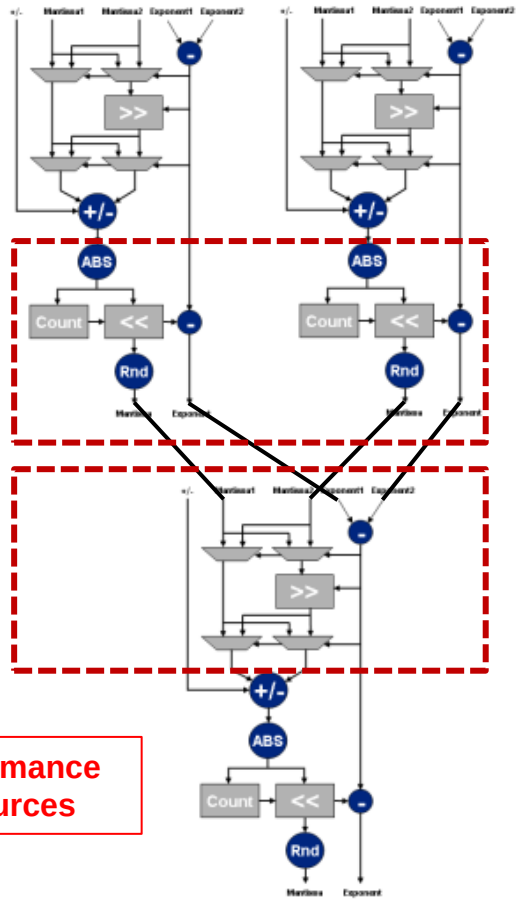
Introducing Fused Datapath

Allows High Performance Floating-Point in FPGAs

© 2011 Altera Corporation—**Confidential**

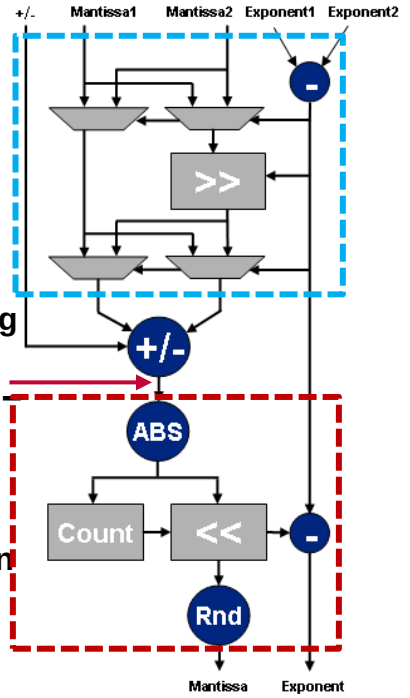
New Floating-Point Implementation

Processor:
Each Operation IEEE754



Low Performance
High Resources

Altera Floating Point:
Fused Datapath

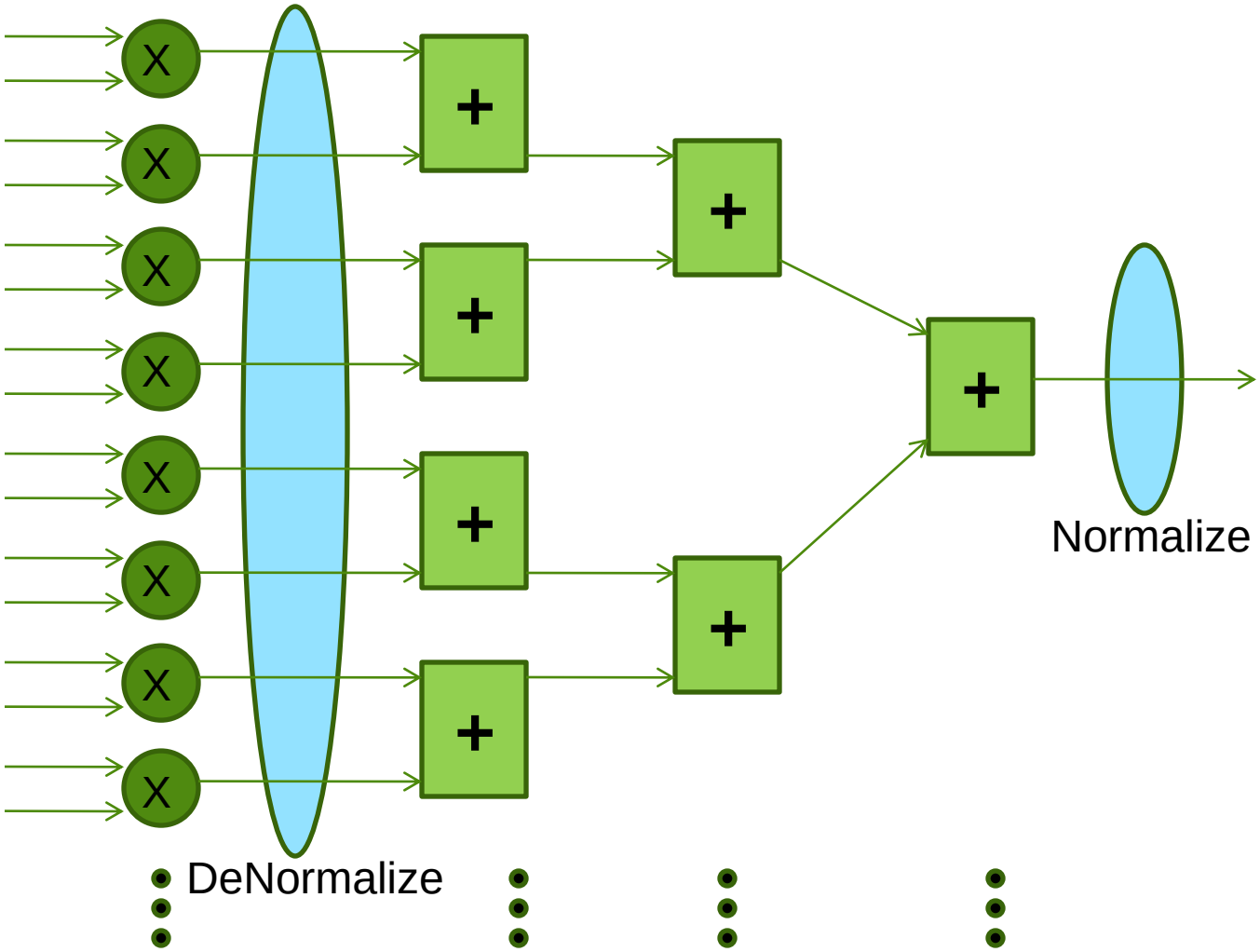


High Performance
Low Resources

© 2011 Altera Corporation—Confidential

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Vector Dot Product Example



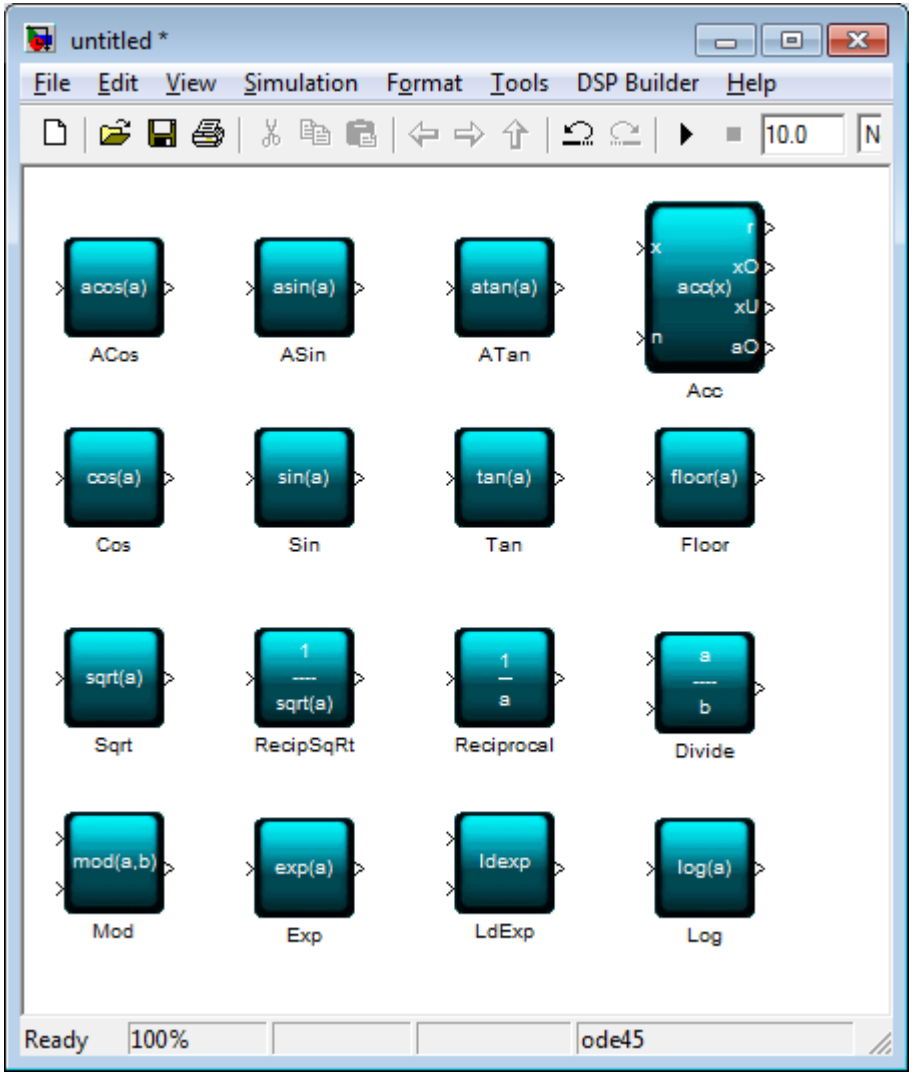
© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Floating Point Functions

- Math.h
 - SIN
 - COS
 - TAN
 - ASIN
 - ACOS
 - ATAN
 - EXP
 - LOG
 - LOG10
 - LDEXP
 - FLOOR
 - CEIL
 - SQRT
 - 1/SQRT
 - DIVIDE
 - MOD

Implemented using "Fused Datapath"



© 2011 Altera Corporation—Confidential

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Stratix V Floating Point Performance Benchmarks

Fast Fourier Transform (FFT)

Matrix Inversion algorithms

- Cholesky Decomposition
- QR Decomposition

© 2011 Altera Corporation—**Confidential**

Altera 28nm high end FPGAs

Stratix V "GS" Family					
Part Number	LEs / ALUTs	ALUTs / Registers	DSP Multiplier Count	Mbits / M20 memory blocks	14 GBps Transceiver Count
5SGSD3	236K	178K / 356K	1200	13 / 688	24
5SGSD4	360K	272K / 543K	2088	19 / 957	36
5SGSD5	457K	345K / 690K	3180	39 / 2014	36
5SGSD6	583K	440K / 880K	3550	45 / 2320	48
5SGSD8	695K	525K / 1050K	3926	50 / 2567	48

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Fast Fourier Transform (FFT) Performance (Mid-size Stratix V, full Floating Point)

FFT MegaCore Device: 5SGSD5	14 Single Precision Floating-point FFT cores, 1,024 pt		
	Usage	Max	%
Logic utilization	317,332	345,200	92%
ALUT	259,844	345,200	76%
Reg	289,781	690,400	42%
Mem bits	1,954,120	41,246,720	5%
M20K	1,190	2,014	59%
18x18 Multipliers	448	3,180	28%
f_{MAX}	**304 MHz**		
Transform time per core	**3.4 us (0.24 us aggregate transform time)**		

28 nm Stratix V FPGA: ~1W per Floating-Point FFT Core

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

FPGA verses DSP Processor

Device	Altera Stratix V 5SGSD8	Texas Instruments TMS320C6678
Resources	695 kLEs 50 Mbits block mem 3926 multipliers 48 TRX (14 GSPS)	8 cores, fixed and SP floating point 1.25 GHz
Peak GMACs (16x16 or 18x18)	**2350** (3926 multipliers @ 600 Mhz)	**320** (40 GMACs per core)
Peak GFLOPs Rating (single precision)	**1000** (see 1 TeraFlop whitepaper)	**160** (20 GFLOPs per core)
1024 length floating point FFT performance (single precision)	**3.41 us** (1024 clock cycles@ 300 MHz)	**10.26 us** (12800 clock cycles @ 1.25 GHz)
Aggregate 1024 length FFT transform time	**0.17 us** (20 FFTs per device)	**1.28 us** (8 FFTs per device, 1 per core)

The Cholesky Decomposition

- The Least Squares solution for x in Ax = b
- A must be Hermitian (conjugate symmetric)
 - Only lower triangular matrix is needed for calculation
- If A is positive definite, it can be decomposed into lower triangular matrix L and conjugate transpose L’ (A= L * L’)
- With Cholesky decomposition, x is solved via forward and backward substitution with decomposed matrices L and L’
- Cholesky decomposition method is more efficient than LU decomposition methods which are suitable for any matrix.

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Solving Diagonal Elements

$$A = \begin{bmatrix} L_{11} & 0 & 0 & 0 \\ L_{21} & L_{22} & 0 & 0 \\ L_{31} & L_{32} & L_{33} & 0 \\ L_{41} & L_{42} & L_{43} & L_{44} \end{bmatrix} \begin{bmatrix} L_{11} & L_{21} & L_{31} & L_{41} \\ 0 & L_{22} & L_{32} & L_{42} \\ 0 & 0 & L_{33} & L_{43} \\ 0 & 0 & 0 & L_{44} \end{bmatrix} = \begin{bmatrix} L_{11}^2 & & & \\ L_{21}L_{11} & L_{21}^2 + L_{22}^2 & & \\ L_{31}L_{11} & L_{31}L_{21} + L_{32}L_{22} & L_{31}^2 + L_{32}^2 + L_{33}^2 & \\ L_{41}L_{11} & L_{41}L_{21} + L_{42}L_{22} & L_{41}L_{31} + L_{42}L_{32} + L_{43}L_{33} & L_{41}^2 + L_{42}^2 + L_{43}^2 + L_{44}^2 \end{bmatrix}$$

ConjugateSymmetric

$$A_{jj} = \sum_{k=1}^{j} L_{jk} * L'_{kj}$$

where j is the column index of the matrix

$$A_{jj} = \sum_{k=1}^{j} L_{jk} * conj(L_{jk})$$

The first non-zero element, at the top of each column can be obtained by:

$$L_{jj} = \sqrt{A_{jj} - \sum_{k=1}^{j-1} L_{jk} * conj(L_{jk})}$$

Equation 1

$$L_{11} = \sqrt{A_{11}}$$

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Off-diagonal Elements

$$A = \begin{bmatrix} L_{11} & 0 & 0 & 0 \\ L_{21} & L_{22} & 0 & 0 \\ L_{31} & L_{32} & L_{33} & 0 \\ L_{41} & L_{42} & L_{43} & L_{44} \end{bmatrix} \begin{bmatrix} L_{11} & L_{21} & L_{31} & L_{41} \\ 0 & L_{22} & L_{32} & L_{42} \\ 0 & 0 & L_{33} & L_{43} \\ 0 & 0 & 0 & L_{44} \end{bmatrix} = \begin{bmatrix} L_{11}^2 & & & ConjugateSymmetric \\ L_{21}L_{11} & L_{21}^2 + L_{22}^2 & & \\ L_{31}L_{11} & L_{31}L_{21} + L_{32}L_{22} & L_{31}^2 + L_{32}^2 + L_{33}^2 & \\ L_{41}L_{11} & L_{41}L_{21} + L_{42}L_{22} & L_{41}L_{31} + L_{42}L_{32} + L_{43}L_{33} & L_{41}^2 + L_{42}^2 + L_{43}^2 + L_{44}^2 \end{bmatrix}$$

$$A_{ij} = \sum_{k=1}^{j} L_{ik} * L'_{kj}$$

where i and j are the row and column indices of the matrix

$$A_{ij} = \sum_{k=1}^{j} L_{ik} * conj(L_{jk})$$

where L_{jk} is the transpose of L_{kj}

Equation 2

$$L_{ij} = \frac{A_{ij} - \sum_{k=1}^{j-1} L_{ik} * conj(L_{jk})}{L_{jj}} \quad \Rightarrow \quad L_{ij} = \frac{A_{ij} - \sum_{k=1}^{j-1} L_{ik} * conj(L_{jk})}{\sqrt{A_{jj} - \sum_{k=1}^{j-1} L_{jk} * conj(L_{jk})}}$$

© 2011 Altera Corporation—Confidential

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Forward Substitution

We now have L and L' thus $A * x = b \rightarrow L * L' * x = b$

If we define: $y = L' * x \rightarrow L * y = b$

L is the lower triangular matrix, y and b are column matrices and b is known in the system so y can be solved by forward substitution

$$y_j = \frac{b_j - \sum_{k=1}^{j-1} y_k * L_{jk}}{L_{jj}}$$

Equation 3

Note that solving for y is very similar to solving for L shown below

$$L_{ij} = \frac{A_{ij} - \sum_{k=1}^{j-1} L_{ik} * conj(L_{jk})}{L_{jj}}$$

Equation 2

Since equations are similar, Cholesky decomposition and forward substitution are combined into the same process. The only difference is that Eq 2 is conjugated

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Backward Substitution

x can be solved by backward substitution, $L' * x = y$

Since L' is an upper triangular matrix, x has to be solved from the bottom to the top, hence why it's called back substitution

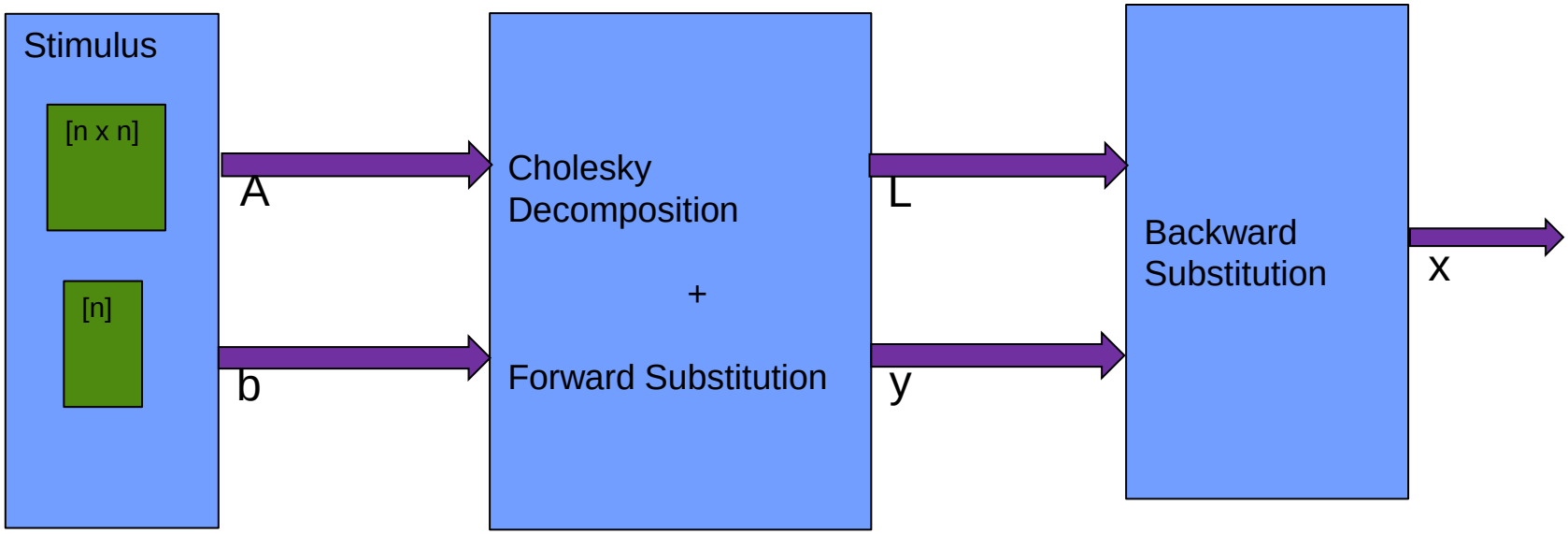
$$x_j = \frac{y_j - \sum_{k=j+1}^{VS} x_k * L'_{jk}}{L'_{jj}}$$

Equation 4

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Cholesky Block Diagram



Solve for x in Ax = b where A is conjugate symmetric

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Performance and FPGA Resources

Cholesky Decomposition Parameterizable Core using 5SGSD5					
Complex Input Matrix Size	Vector Size	ALUTs / Memory blocks / 27x27s	% ALUTs / % Memory blocks / % 27x27s	Latency @ Operating frequency	GFLOPS per core (complex single precision)
30x30	30	76.5K 793 M20K 146 DSP	22% 39% 9%	255 us @ 250 MHz	21.7
60x60	60	141K 955 M20K 268 DSP	41% 47% 17%	328 us @ 235 MHz	39.0
240x240	60	154K 1820 M20K 268 DSP	45% 90% 17%	922 us @ 220 MHz	74.2
360x360	90	204K 1411 M20K 391 DSP	59% 70% 25%	1103 us @ 190 MHz	91.8
400x400	100	220K 1619 M20K 430 DSP	64% 80% 27%	1342 us @ 190 MHz	103

GFLOPs and GFLOPs/Watt

Cholesky Decomposition Parameterizable Core using 5SGSD5					
Complex Input Matrix Size (n x n)	Vector Size	Through-put (Matrix per second)	GFLOPS per core (complex single precision)	Core power consumption as measured using Altera 5SGSD5 eval board	GFLOPs/Watt
30x30	30	472,464	21.7	7.7 W	2.8
60x60	60	118,858	39.0	13.6 W	2.9
240x240	60	8,467	74.2	14.0 W	5.3
360x360	90	1142	91.8	14.7 W	6.2
400x400	100	1182	103	16.1 W	6.4
Complex Cholesky FLOPs = $4/3n^3 + 8n^2$					

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Competive Results: Nvidia GPU

Cholesky Decomposition (single precision)			
Matrix Size	GFLOPs with LAPACK Library	GFLOPs with Magma Library	GFLOPs with Nvidia OpenCL Library
512x512	20	22	58
768x768	20	39	82
1024x1024	36	57	68
2048x2048	60	117	96
Cholesky FLOPs = 4 $N^3/3$, where N is matrix dimension			

- Results in about 0.25 GFLOPs/Watt (512x512)
- Nividia GTX480 rated at 977 GFLOPs
- Intel Pentium4 3.7GHz rated at 14.8 GFLOPs

High Performance Relevance Vector Machine on GPUs
Depeng Yang, Getao Liang, David Jenkins,Gregory D. Peterson, and Husheng Li
U of Tennessee, Knoxville

© 2011 Altera Corporation—Confidential

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

More Nvidia Results

LU Decomposition (single precision)			
Matrix Size	CPU GFLOPs	GPU GFLOPs	GPU speedup
1024x1024	24.2	51.4	3.1
2048x2048	26.5	111.7	5.2
3072x3072	27.5	151.6	6.5
4032x4032	29.96	183.02	7.1
Using Magma 1.0 RC5 library			

- Nvidia Fermi Tesla C2050, 1147.0 MHz clock
- AMD Quadro NVS 290, 918.0 MHz clock

MAGMA LAPACK for GPUs
Stan Tomov, Research Director, Innovative Computing Laboratory
Department of Computer Science
University of Tennessee, Knoxville

© 2011 Altera Corporation—Confidential

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

QR Decomposition

- QR Solver finds solution for Ax=b linear equation system using QR decomposition, where Q is ortho-normal and R is upper-triangular matrix. A can be rectangular.

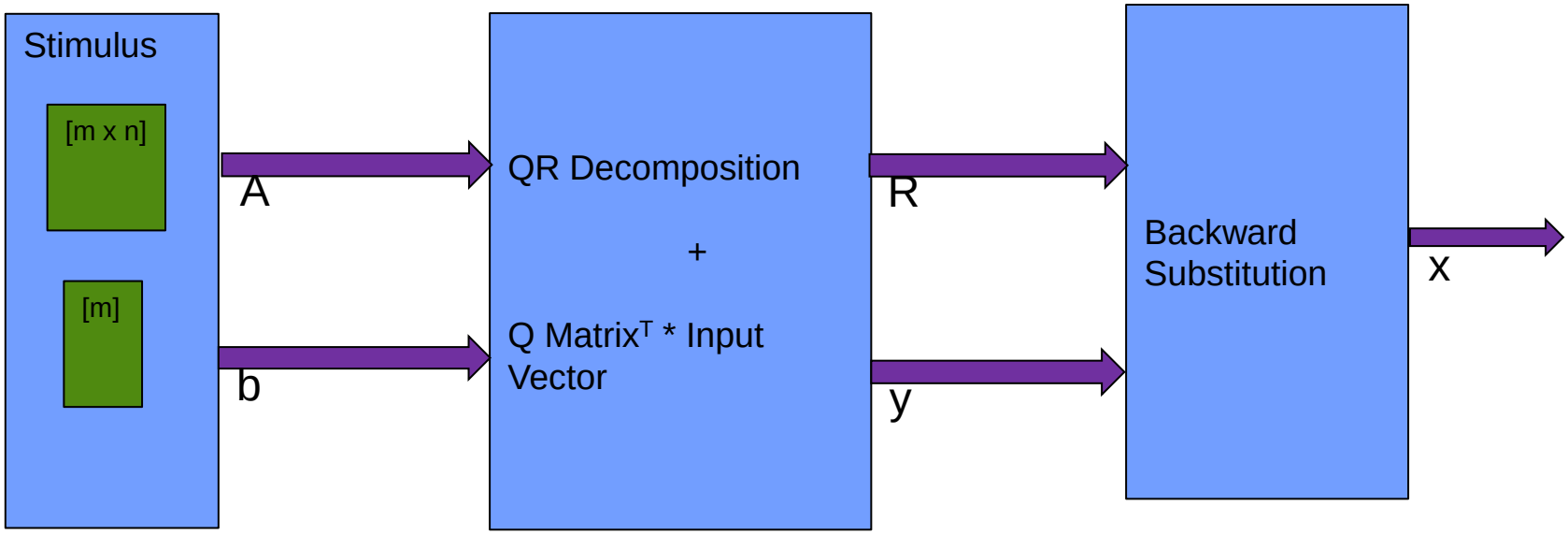
- Steps of Solver
 - *Decomposition:* $A = Q \cdot R$
 - *Ortho-normal property:* $Q^T \cdot Q = I$
 - *Substitute then mult by Q^T:* $Q \cdot R \cdot x = b$ $R \cdot x = Q^T \cdot b = y$
 - *Backward Substitution:* $Q^T \cdot b = y$ *solve* $R \cdot x = y$

- *Decomposition is done using Gram-Schmidt derived algorithms. Most of computational effort is in "dot-product"*

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Block Diagram



Solve for x in Ax = b where A is non-symmetric, may be rectangular

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Performance and FPGA Resources

QR Decomposition Parameterizable Core using 5SGSD5					
Complex Input Matrix Size	Vector Size	ALUTs / Memory blocks / 27x27s	% ALUTs / % Memory blocks / % 27x27s	Latency @ Operating frequency	GFLOPS per core (complex single precision)
50x100	50	105K 230 M20K 227 DSP	30% 11% 14%	45 us @ 250 MHz	43.8
100x200	50	106K 304 M20K 228 DSP	31% 15% 14%	213 us @ 250 MHz	64.3
100x200	100	202K 504 M20K 428 DSP	58% 25% 27%	173 us @ 200 MHz	91.9
250x400	100	200K 858 M20K 428 DSP	58% 43% 27%	1586 us @ 200 MHz	106
400x400	100	203K 1566 M20K 428 DSP	59% 78% 27%	4029 us @ 200 MHz	106

GFLOPs and GFLOPs/Watt

QR Decomposition Parameterizable Core using 5SGSD5					
Complex Input Matrix Size (n x m)	Vector Size	Through-put (Matrix per second)	GFLOPS per core (complex single precision)	Core power consumption as measured using Altera 5SGSD5 eval board	GFLOPs/Watt
50x100	50	31,681	43.8	10.8 W	4.1
100x200	50	5,920	64.3	13.9 W	4.6
100x200	100	8,467	91.9	21.0 W	4.4
400x400	100	310	106	25.2 W	4.2
450x450	75	165	80.0	20.2	4.0
Complex QRD FLOPs = $5.33mn^2 + 8mn - 2n + 4n^2$					

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Accuracy, Validation, and summary

© 2011 Altera Corporation—**Confidential**

Computational error analysis

QR Decomposition Accuracy			
Complex Input Matrix Size (n x m)	Vector Size	MATLAB using computer Norm/Max	DSPBA generated RTL Norm/Max
50x100	50	5.01e-5 / 6.42e-6	4.87e-5 / 6.02e-6
100x200	100	2.3e-5 / 1.24e-6	1.68e-5 / 9.97e-7
400x400	100	8.8e-5 / 4.81e-6	7.07e-5 / 4.03e-6
using Frobenius norm $\|E\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\left\|e_{ij}\right\|^2}$			

Cholesky Decomposition results are similar

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

Summary

- High performance floating point designs can be built using FPGAs
 - High density of 27x27, 36x36, 54x54, 72x72 multipliers available at 28nm
 - New floating point toolflow reduces routing density to sustainable level
 - Availability of optimized math.h library of floating point functions
- FPGA Fixed point parellelism performance benefits now carry over into floating point
- Best in class GFLOPs / Watts
- Real-world, not marketing, floating point benchmarks for comparison

© 2011 Altera Corporation—**Confidential**

ALTERA, ARRIA, CYCLONE, HARDCOPY, MAX, MEGACORE, NIOS, QUARTUS & STRATIX are Reg. U.S. Pat. & Tm. Off. and Altera marks in and outside the U.S.

An IA-32 Processor with a Wide Voltage Operating Range in 32nm CMOS

Gregory Ruhl, Saurabh Dighe, Shailendra Jain, Surhud Khare, Satish Yada, Ambili V, Praveen Salihundam, Shiva Ramani, Sriram Muthukumar, Srinivasan M, Arun Kumar, Shasi Kumar, Rajaraman Ramanarayanan, Vasantha Erraguntla, Jason Howard, Sriram Vangal, Paolo Aseron, Howard Wilson, Nitin Borkar

Microprocessor & Programming Research, Intel Labs

Intel Confidential

Purpose

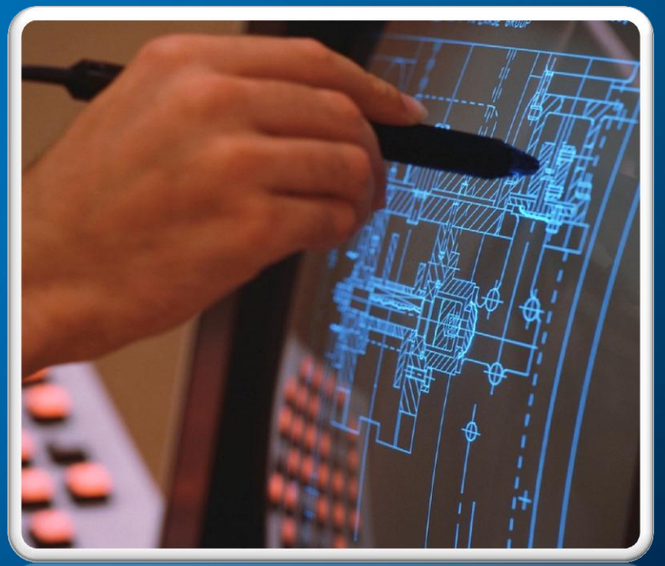
Claremont: Near Threshold Voltage and Wide Dynamic Range IA Core



Demonstrate energy benefits of Near Threshold Voltage (NTV) computing to IA



Extend dynamic range of operation from NTV to Vmax for energy efficient performance



Advance low voltage, variation aware and multi-corner design methodologies

Agenda

- Design Challenges
- Claremont Prototype
- Design Strategies and Methodologies
 - NTV Design
 - Wide Dynamic Range Design
- Results and Summary

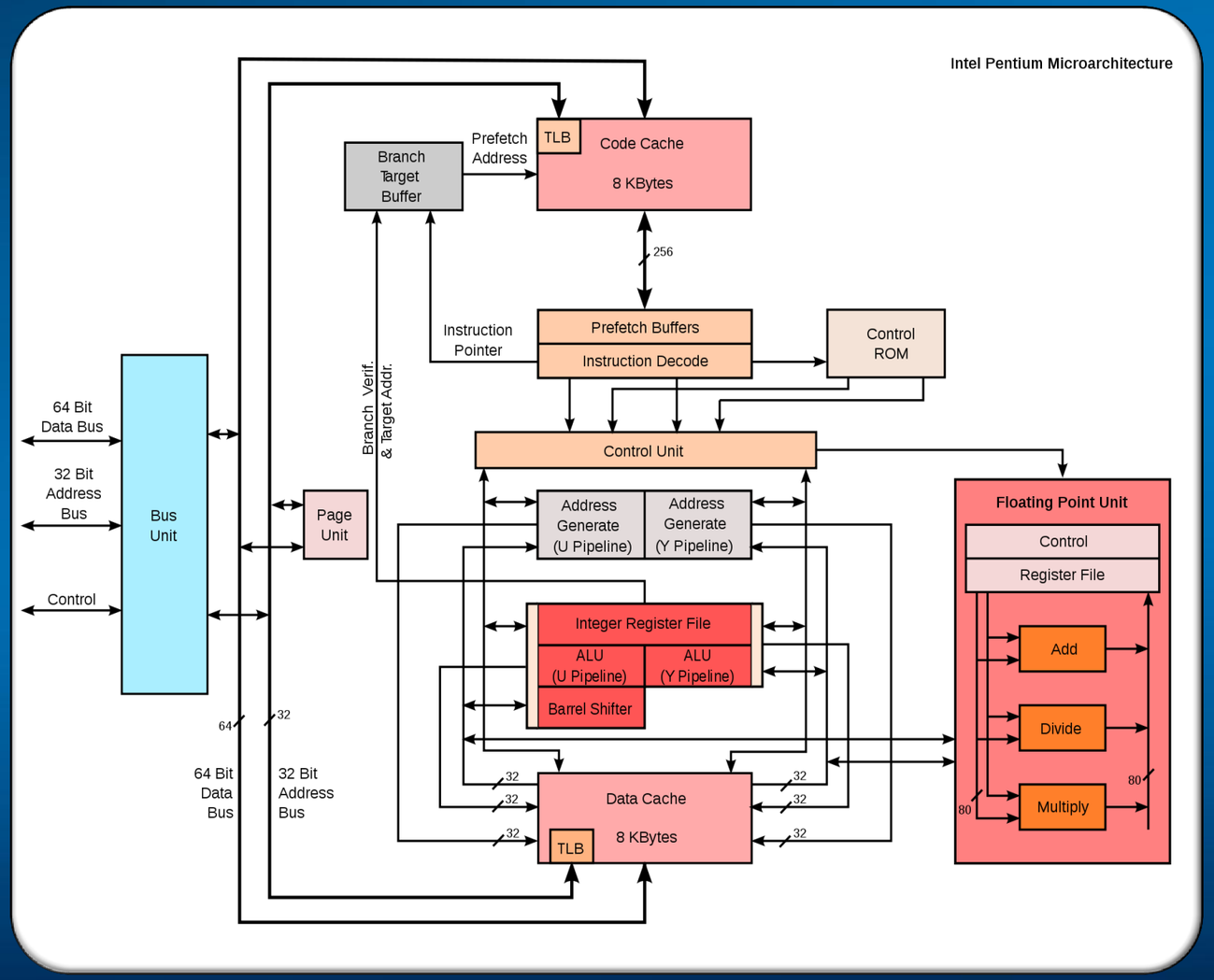
Design Challenges

- Reduced I_{on}/I_{off}, noise margins and variability results in circuit functional failures
- Power/performance profile becomes extremely sensitive to PVT variations
- Tools and methodologies are not mature for low voltage designs
- Wide dynamic range design convergence is complicated by
 - Disproportionate device vs. interconnect delay scaling
 - Multiple voltage domains

Agenda

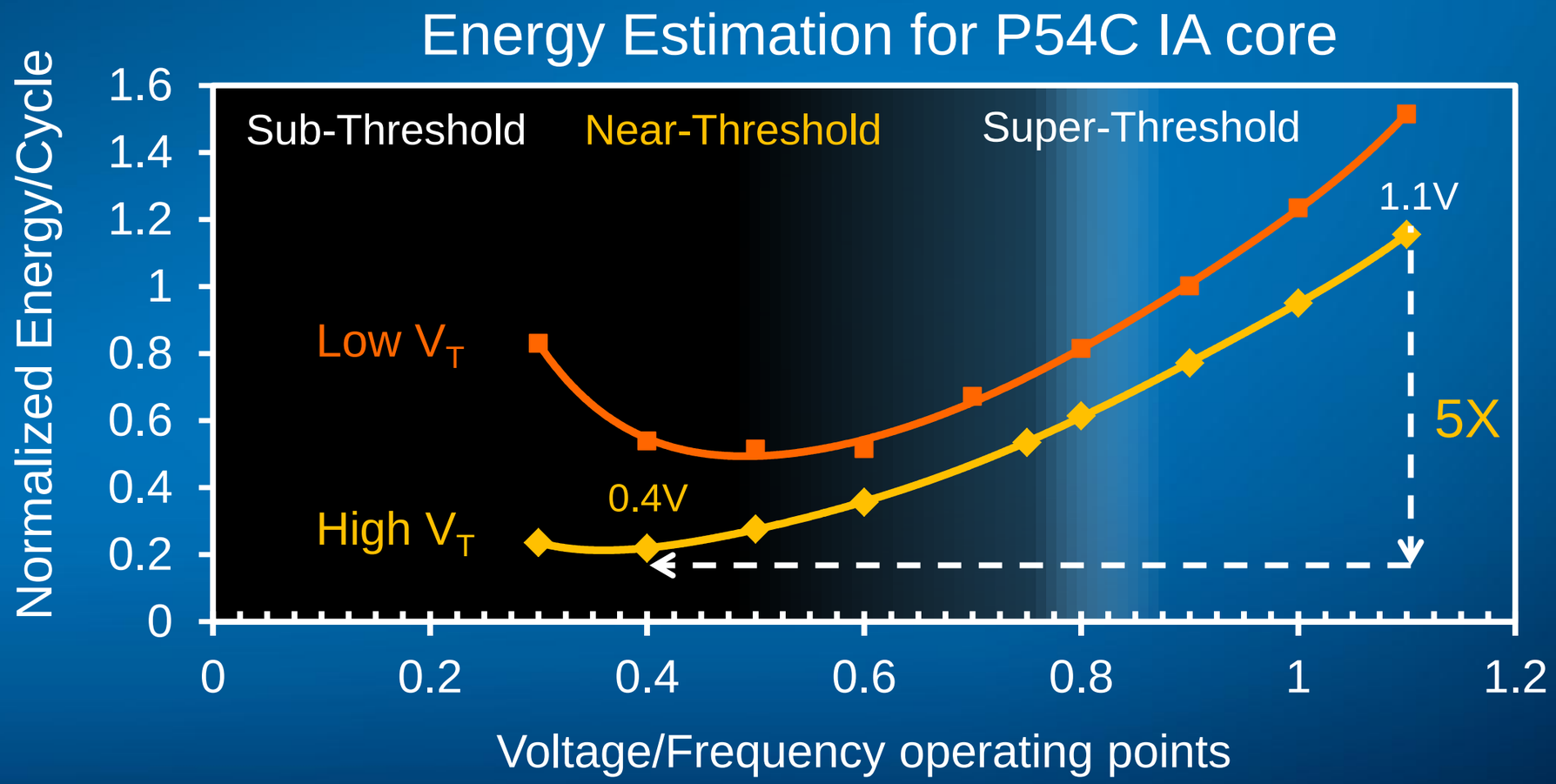
- Design Challenges
- Claremont Prototype
- Design Strategies and Methodologies
 - NTV Design
 - Wide Dynamic Range Design
- Results and Summary

P54C IA-32 Core Background



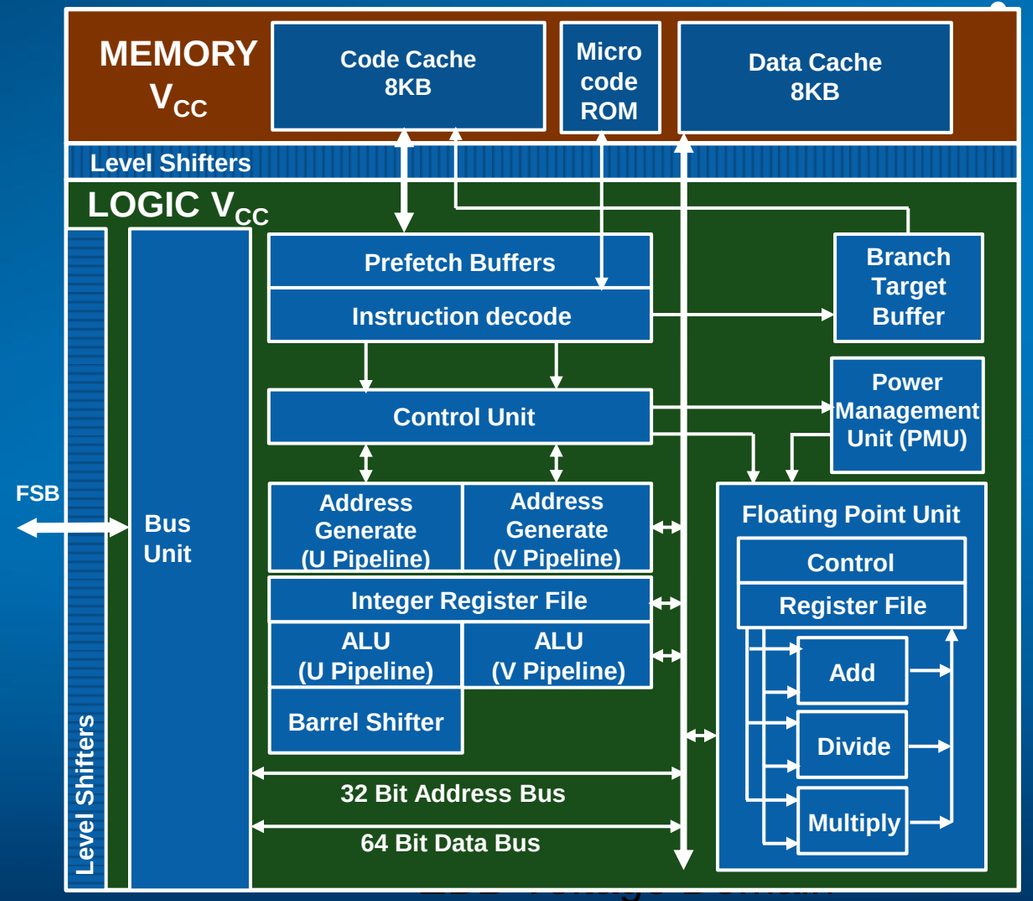
- Legacy Pentium ® core (1994)
- 32-bit CPU with 64-bit data bus
- Superscalar, in-order pipeline architecture with pipelined floating point unit
- Dynamic branch prediction
- Separate code and data caches (8KB)
- Fractional bus operation allowing core frequencies higher than 66MHz FSB

Setting the Design Targets



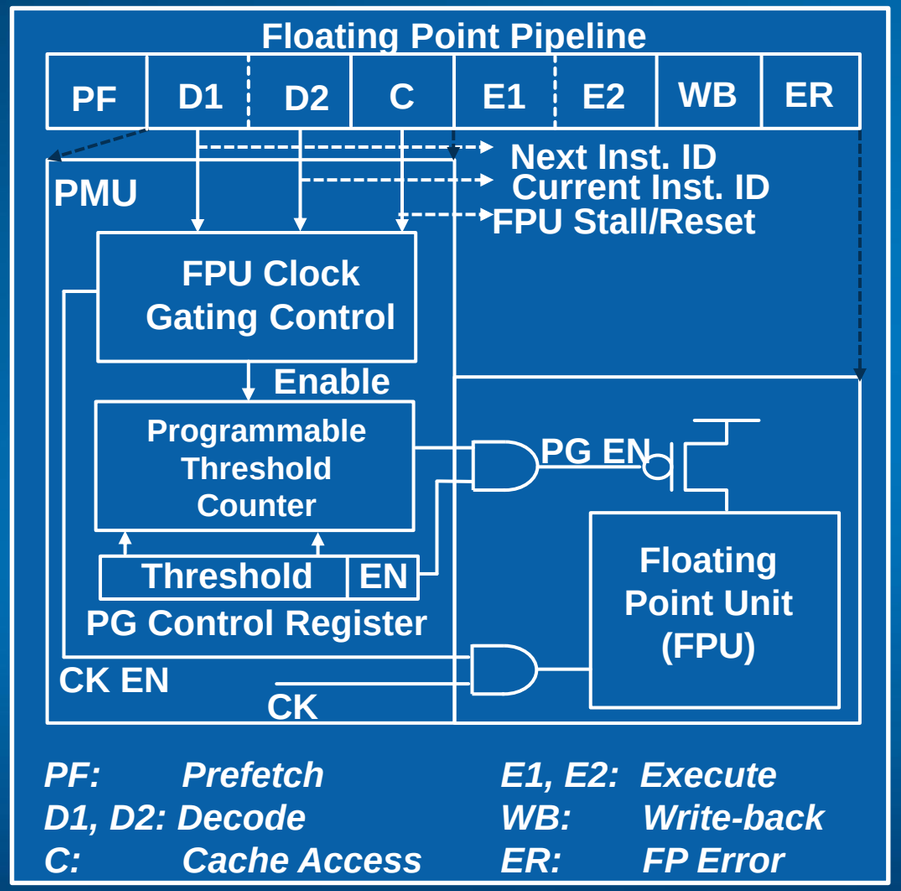
~5X efficiency improvement with aggressive voltage scaling

Claremont Prototype



- P54C IA Core in 32nm CMOS
- Aggressive Voltage Scaling
 - Vmin Target: 0.5V (Logic) and 0.55V (RF)
 - Low Voltage, Variation Aware Design
- Ultra Low Power Design
 - Sub-20mW Total Core Power Target at 0.5V
- Wide Dynamic Operating Range
 - Triple Corner Optimization
 - 0.5V/66MHz, 0.75V/333MHz, 1.05V/525MHz

Proactive Power Management



- **Instruction-driven power gating**
 - Dynamically turn-off FPU during idle periods
 - 65% FPU sleep enabled
 - Single cycle wake-up
- **Programmable sleep-threshold**
 - Based on application and operating point
 - Energy saved **>** Wake-up overheads

Workload-aware, fine-grain power management

Agenda

- Design Challenges
- Claremont Prototype
- Design Strategies and Methodologies
 - NTV Design
 - Wide Dynamic Range Design
- Results and Summary

Near Threshold Voltage Logic

- Variation aware library pruning to ensure reliable NTV operation
- Limited transistor stacks to 3, No wide TG muxes, No contention circuits
- Pruned minimum sized and low drive strength cells
- Sequentials redesign with interruptible and upsized keepers
- 10T single ended transmission gate register file cell topology
- Semi interruptible split output level shifters
- Full swing 3.3V I/O for legacy board compatibility

Low Voltage Timing Convergence

Random/Systematic process variations ➔ Path delay uncertainty ➔ Max/min failures

- Max convergence strategy
 - Setup violations are not fatal; Can be corrected by relaxing PV phase
 - Conservative library pruning ➔ Low variation impact ➔ Better PV to silicon correlation
 - Shallow P54C pipeline (~75 gate stages) ➔ Averaging effect of random variations
- Min convergence strategy
 - Hold violations are fatal; Important to ensure sufficient hold margins
 - Tango charges data/clock path variation; Need to consider sequential hold variation as well..

“Variation aware” timing convergence is essential

Hold Margin Guard Banding

Hold time variation characterization using NOVA/MPP2

Sequential Variants	Hold Time TZST, 20C		
	0.5V, 0σ	0.5V, 5.5σ	0.4V, 5.5σ
✗ 1X Local Clock Inverters	-159 pS	686 pS	9899 pS
✓ 2X Local Clock Inverters	-224 pS	136 pS	786 pS

Variation aware hold margin guard banding for robust sub-0.5V operation

Agenda

- Design Challenges
- Claremont Prototype
- Design Strategies and Methodologies
 - NTV Design
 - Wide Dynamic Range Design
- Results and Summary

Wide Dynamic Range Design Challenge

P1269.4	Critical Path at 0.5V	Critical Path at 1.05V
Device Delay (DD) Contribution	98%	75%
Interconnect Delay (ICD) Contribution	2%	25%
Characteristics	Device dominated data paths become most critical. Effect of ICD is negligible	ICD becomes significant component at higher voltages

The change in critical path distribution across the wide voltage range significantly increases timing optimization efforts to achieve targeted frequency at a given voltage corner

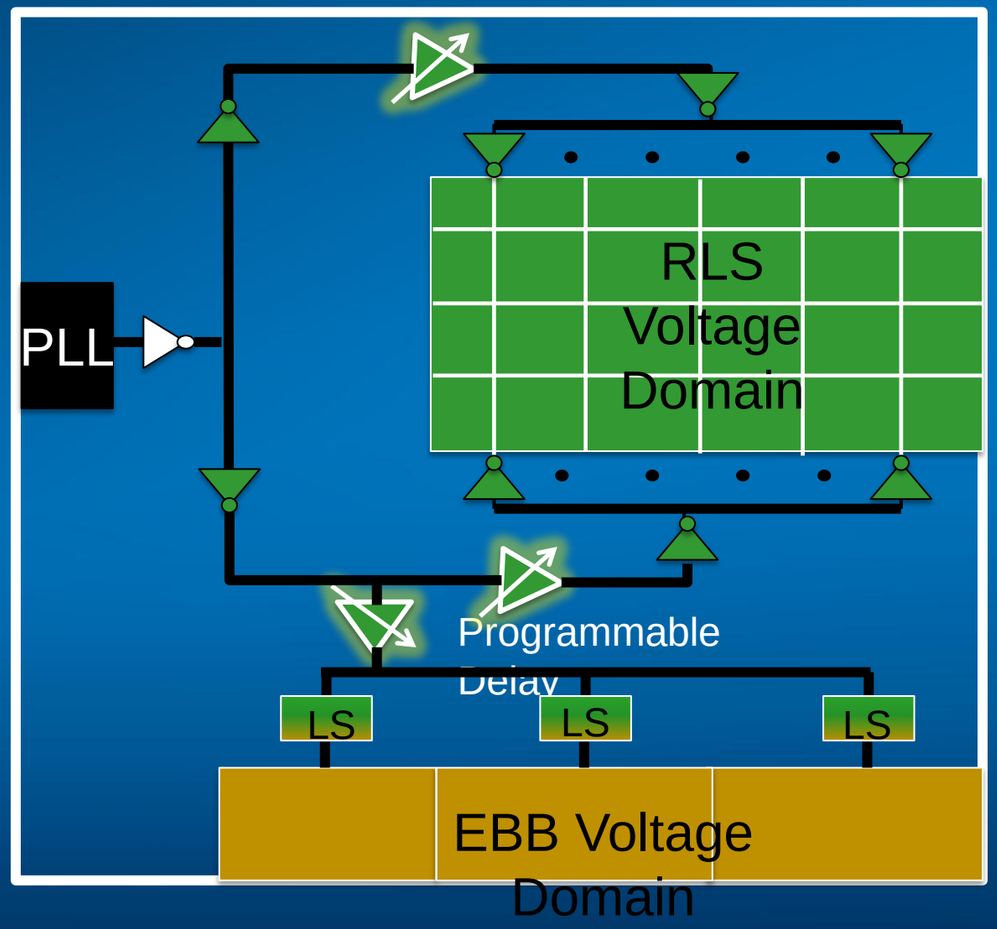
Wide Dynamic Range Design Optimizations

Synthesis Corner Evaluation	Timing Metrics					
	0.75V			0.5V		
	WNS	TNS	Fmax	WNS	TNS	Fmax
✗ 0.75V, 450MHz	-0.18 nS	-165 nS	420 MHz	-1.8 nS	-1950 nS	74 MHz
✓ 0.5V, 86MHz	-0.16 nS	-21 nS	430 MHz	-0.3 nS	-4.6 nS	84 MHz

- Insignificant ICD at ULV ➔ Suboptimal P&R ➔ ICD dominated critical paths at high voltages
 - Prioritized ICD dominated paths at partition level before placement
- Synthesis constraints: Superset of timing constraints required across voltage range

Multi corner constraints for single corner synthesis

Multi-Voltage Clocking and Skew Management



- Core logic (RLS) & memories (EBB) operate on independent power supply
- Level shifters in clock distribution network
- Inter-block skew is voltage dependent
- Programmable delay buffers for the skew management, configured via scan
- 1.5-2X skew reduction across different RLS, EBB voltage combinations

Enhanced PV Methodology

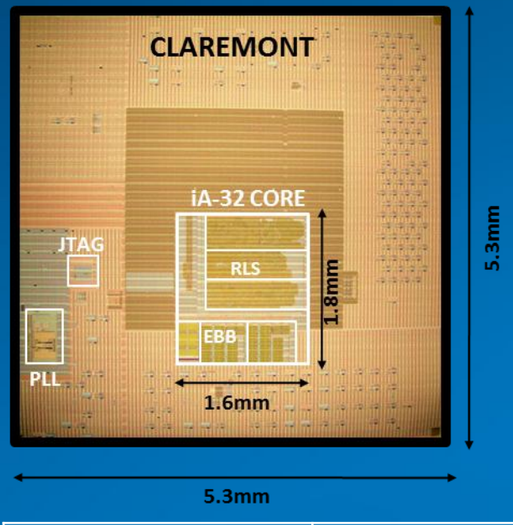


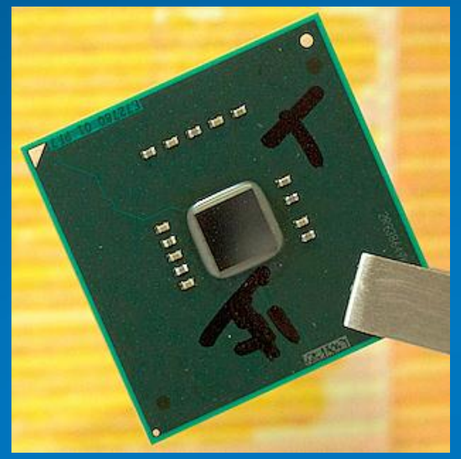
- Modified intra-block & inter-block skew function in Tango for multi-voltage timing analysis and roll-ups
- Incorporated block Voltage Mapping (VM) table and Voltage Dependent Skew (VDS) table in Tango environment
- Skew is computed based on operating voltage of launching and sampling block using entries in VM & VDS tables

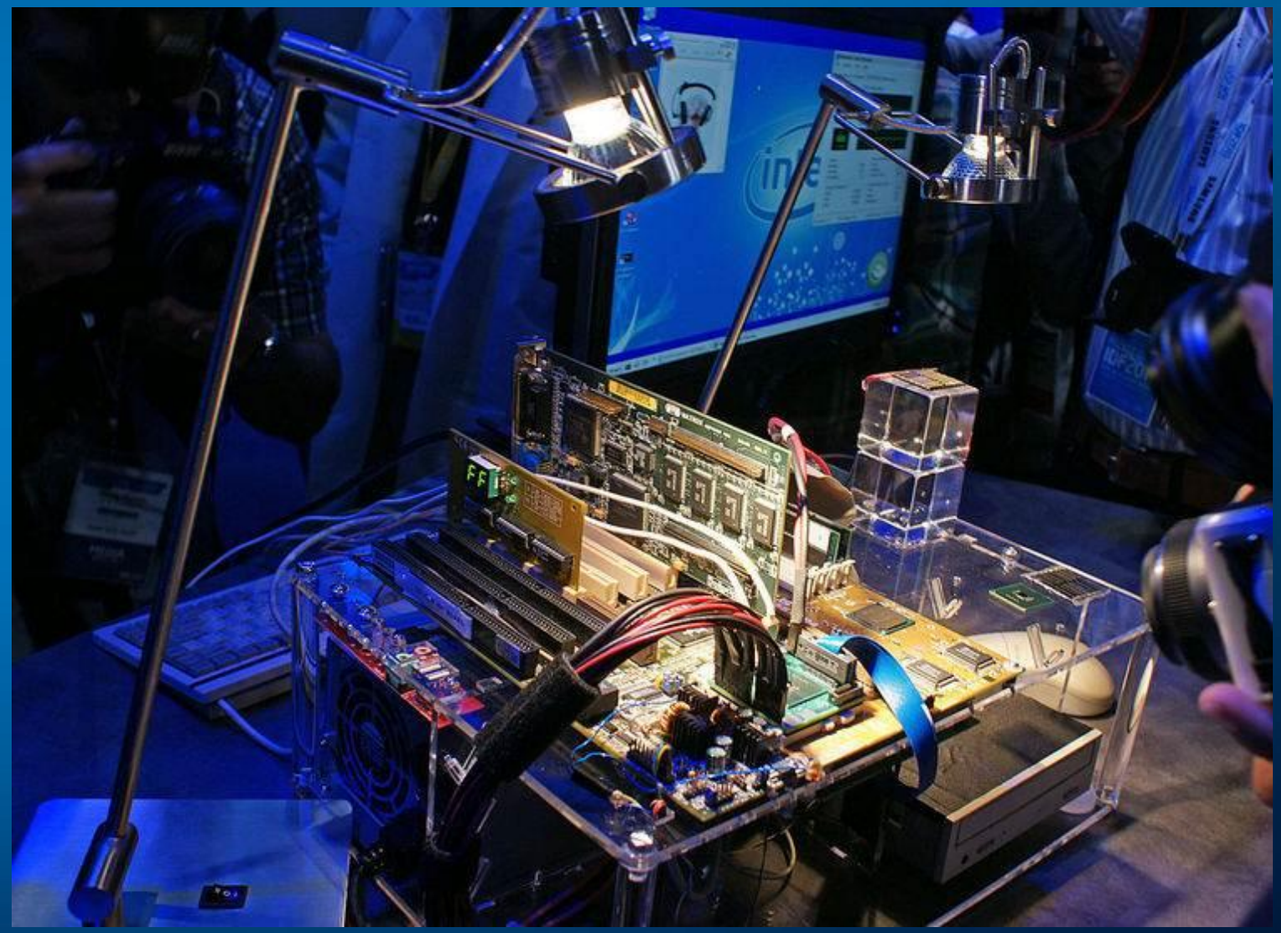
Agenda

- Design Challenges
- Claremont Prototype
- Design Strategies and Methodologies
 - NTV Design
 - Wide Dynamic Range Design
- Results and Summary

Claremont Die Micrograph and Test Setup

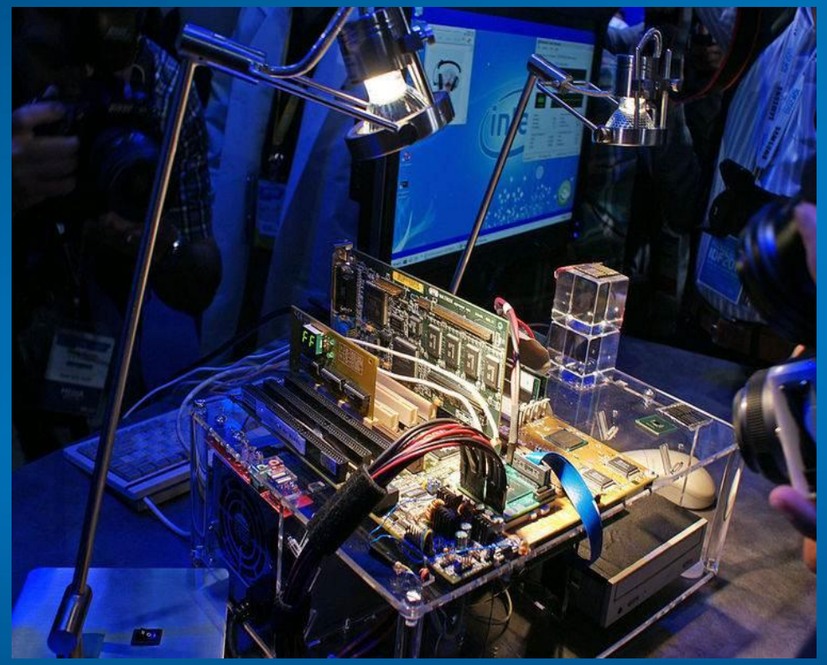






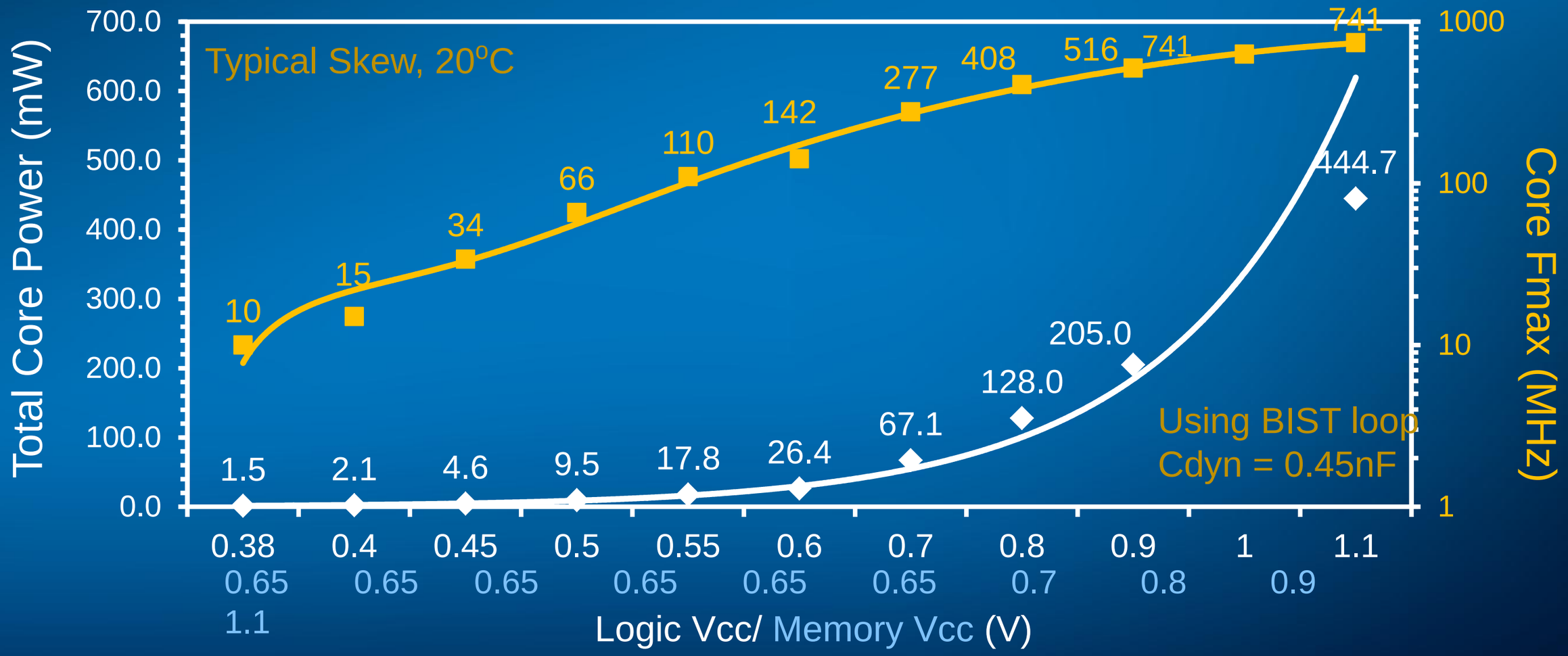
Technology	32nm High-K Metal Gate Technology
Interconnect	1 Poly, 9 Metal (Cu)
Transistors	6 Million (core & EBBs)
Core Area	1.96mm^2
Signals	168
Package	951 Pins FCBGA11

Claremont Test Challenges



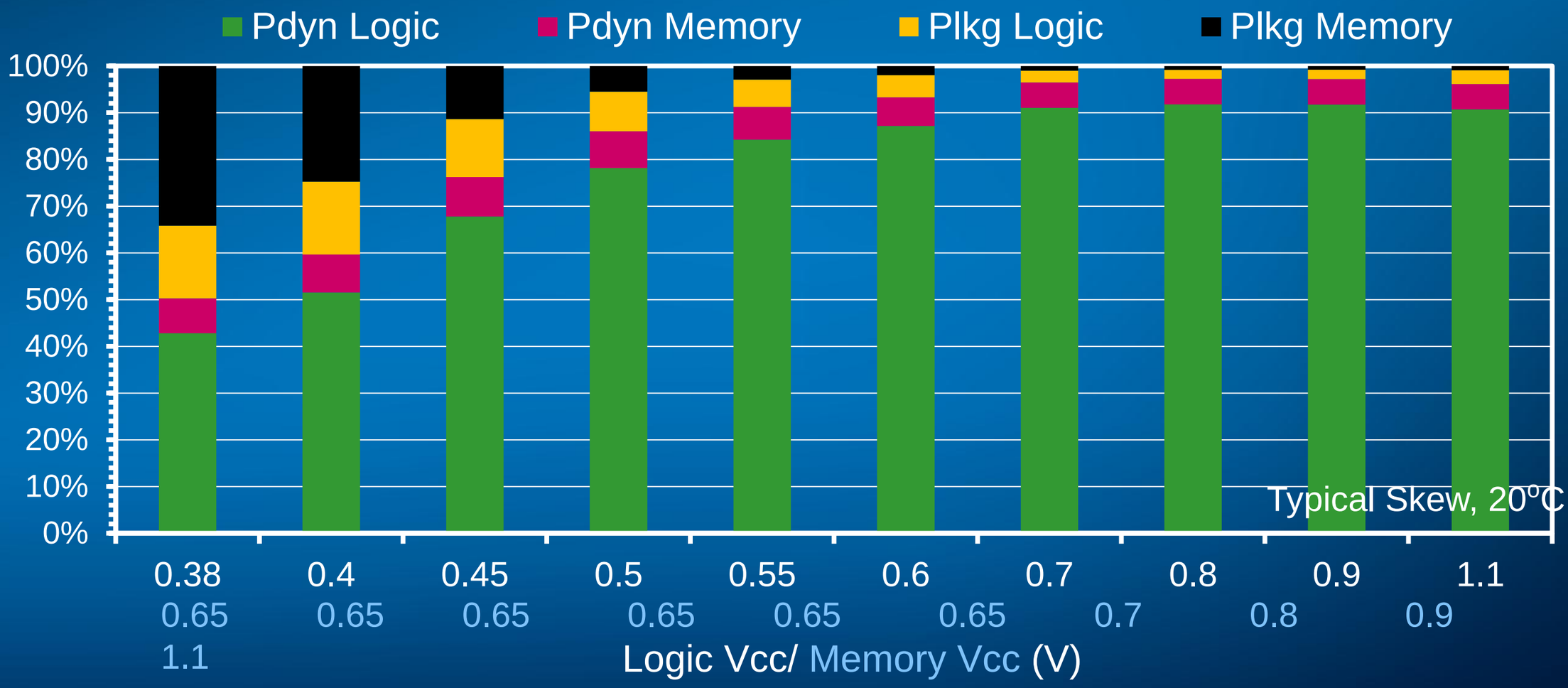
- 18 year old motherboard
 - Age variation
- Most peripherals fail below 15Mhz
- Front side bus spec timing and voltage challenges
- Lack of uBreak points & advanced debug features

Power/Performance Characteristics



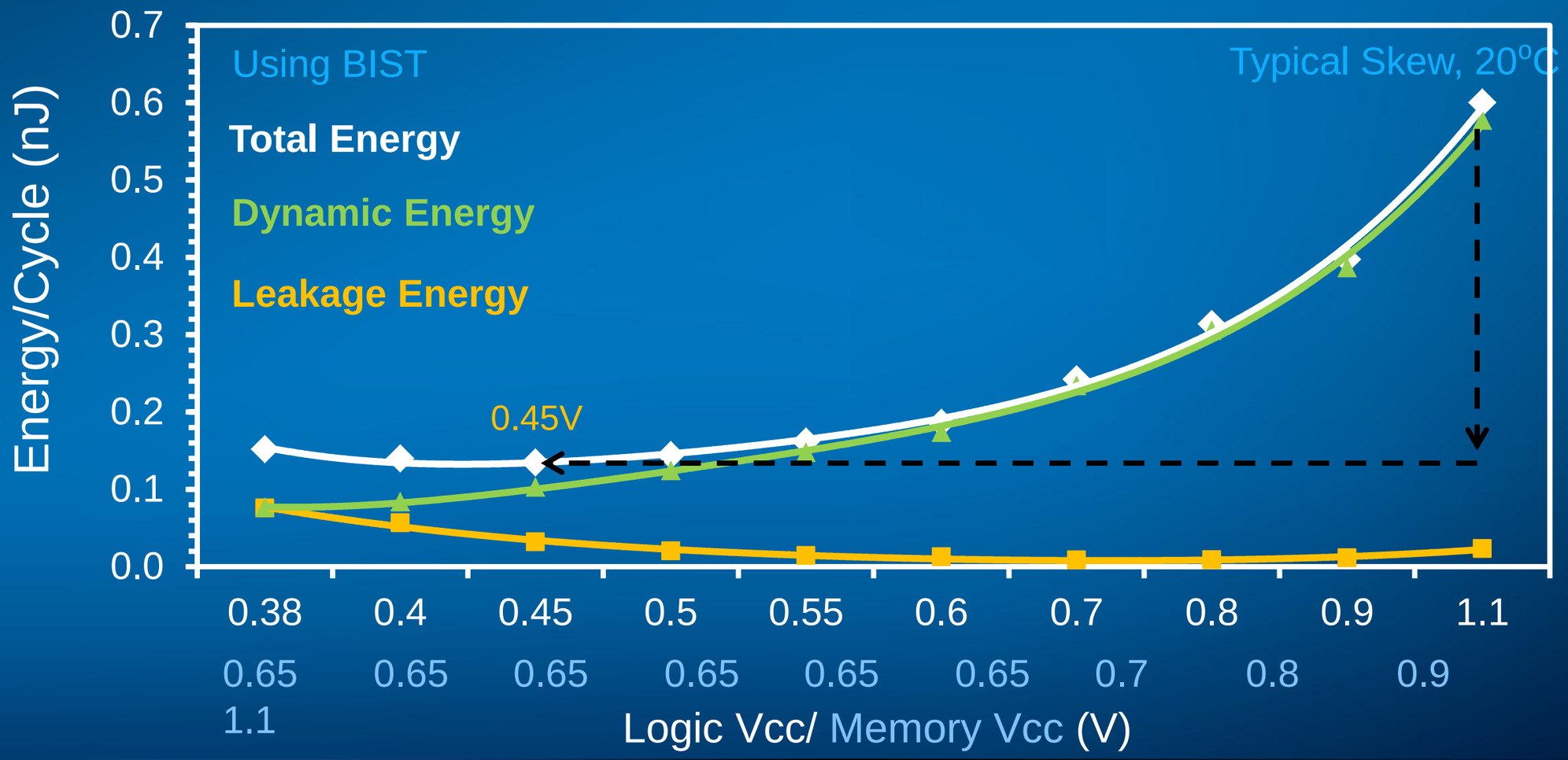
Wide Dynamic Range 1.1V/741MHz/445mW to 380mV/10MHz/1.5mW

Total Power Breakdown



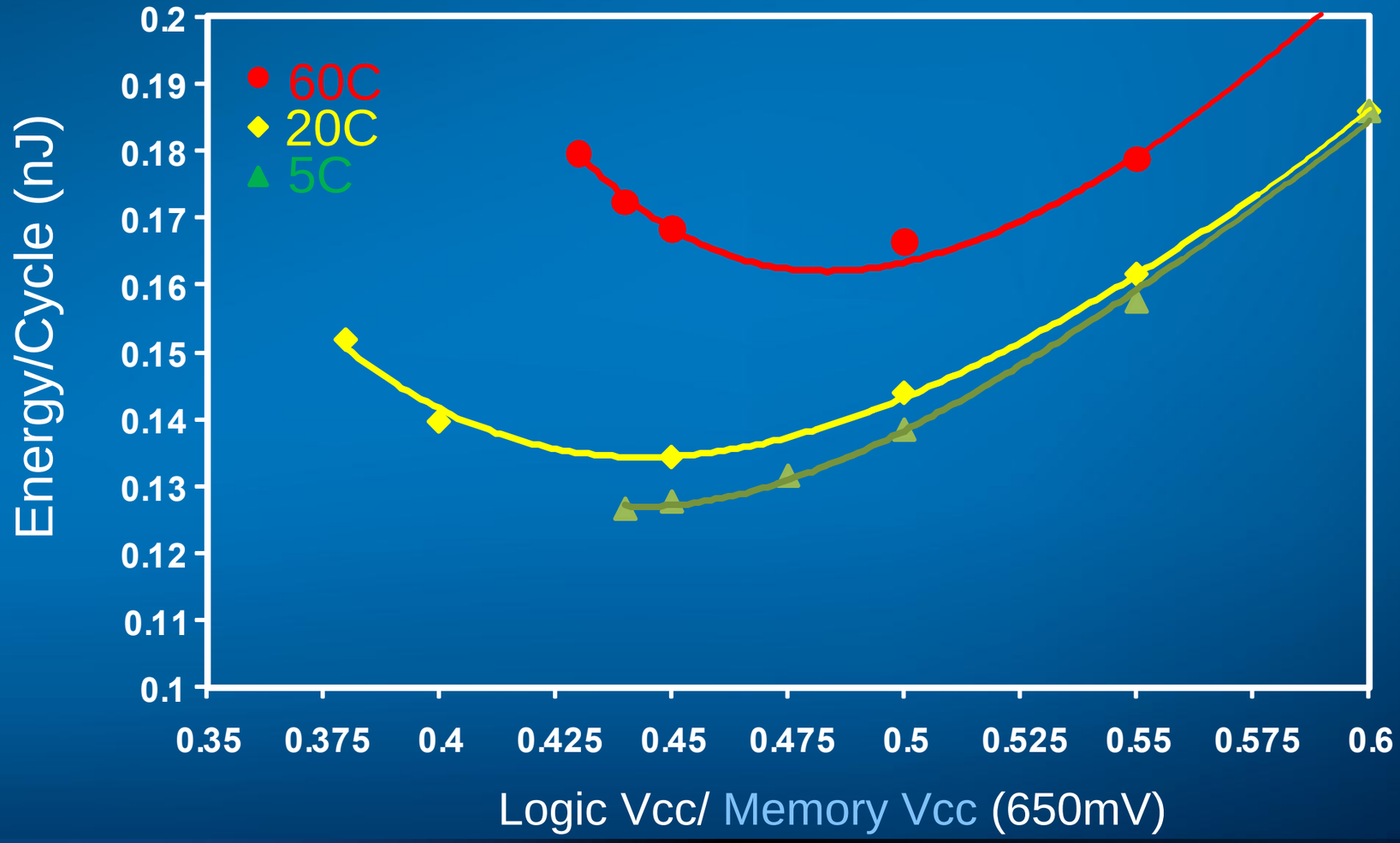
Leakage power scales from 3% @1.1V to 50% @ 0.38V

Energy/Cycle: Typical Skew



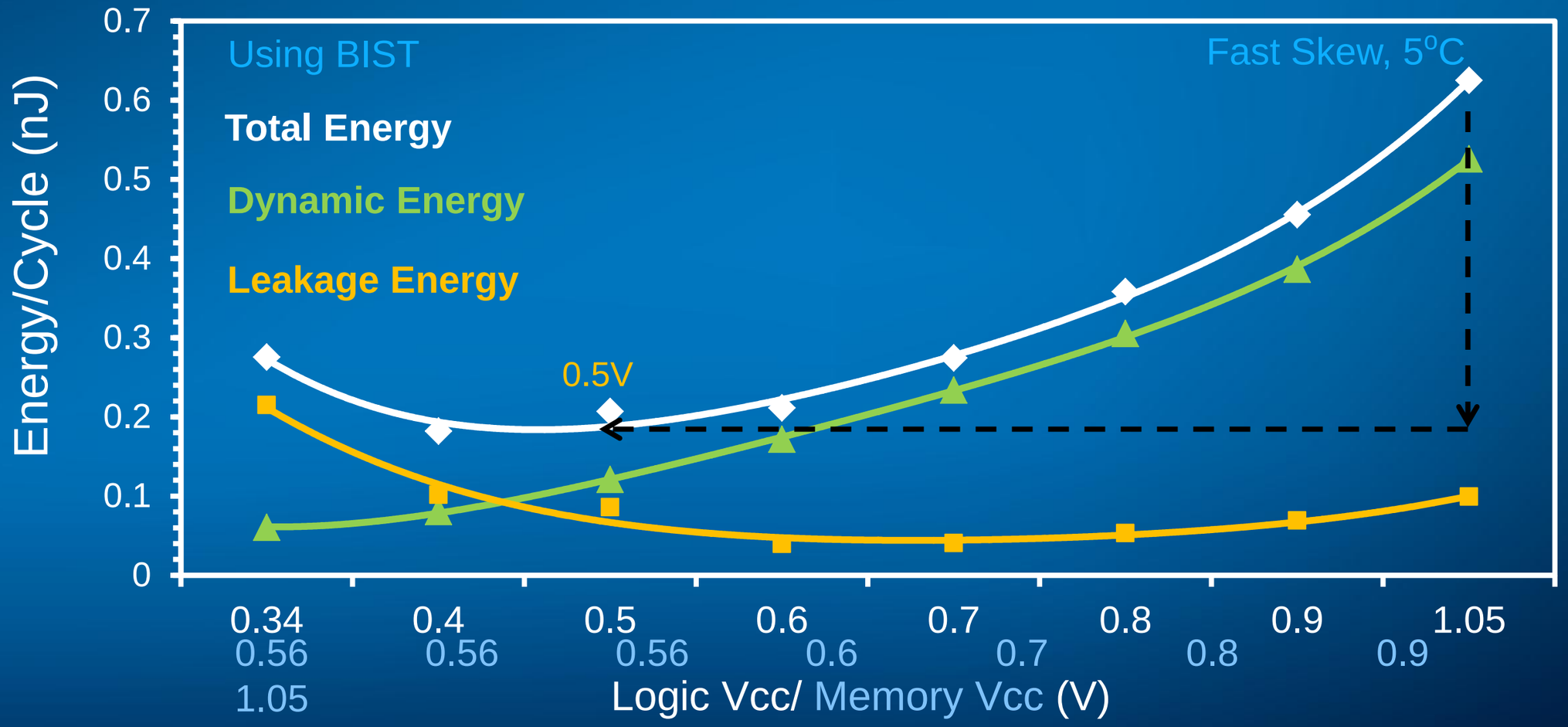
4.5X Energy reduction from Vmax: 135pJ/cycle at 450mV

Temperature Dependence: Typical Skew



25% increase from 5C to 60C

Energy/Cycle: Fast Skew



3X Energy reduction from Vmax: 200pJ/cycle at 500mV

Area Penalty per Technique

Technique	Increase from Audit DB
Modified Sequentials	27%
0.5V 66MHz Target	24%
Complex cells Pruning	10%
3 RLS blocks vs. 1	8%
Min Z	5%
Additional Min fixing	2%

- **Area overhead is a non-linear function of Vccmin improvement.**
- **Incremental Vccmin improvement is more practical and will have a lower penalty**

Claremont: Industry's First NTV IA Core

380mV

Core Logic Vmin

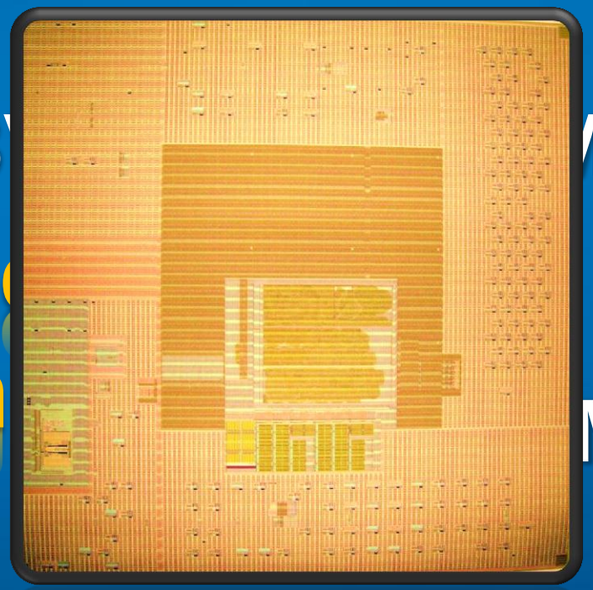
1.5m
W

Total Core Power at Vmin

0.38

Wid
Ran

1.1V,



MHz

MHz

293X

Total Power Reduction from Vmax to Vmin

4.5X

Total Energy Savings from Vmax to Vopt

A0 silicon booting Multiple O/S

Measured using BIST

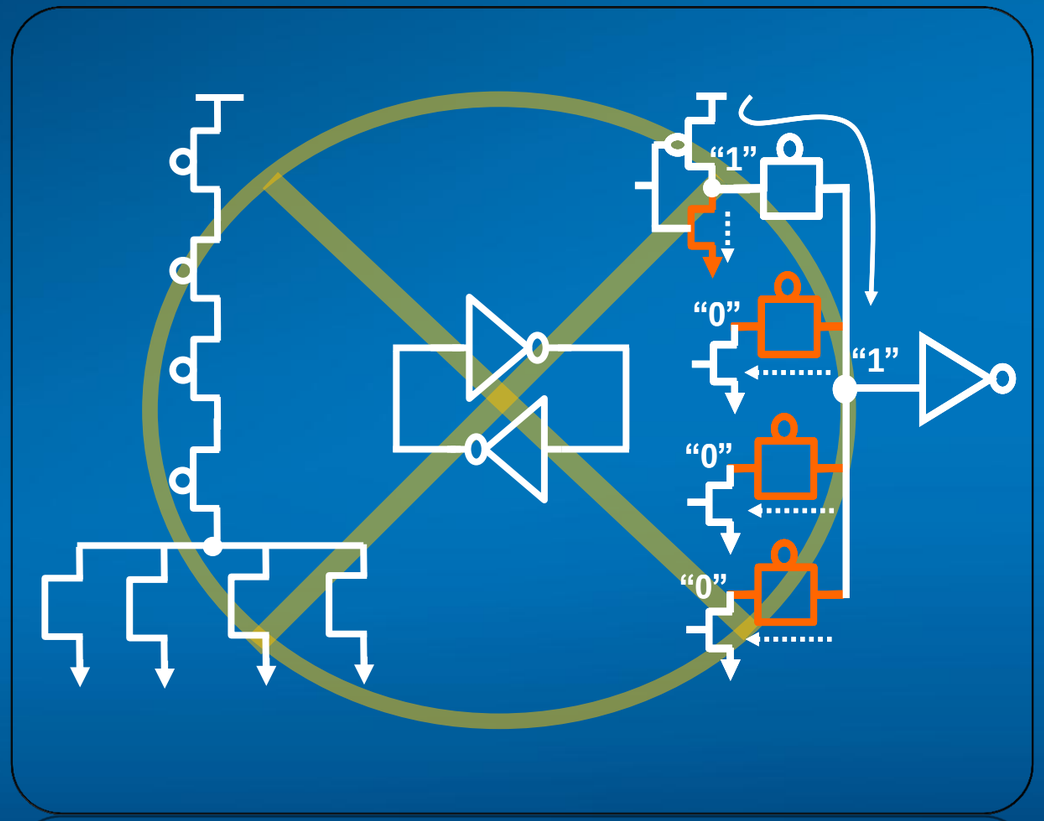
Conclusion

- NTV: A Promising Technology for Energy Efficient Computing
 - Beneficial for ultra-low power IA with modest performance demands
 - Energy efficient SOCs, Graphics, Sensor hubs, Many-core CPUs, Exascale…
- Claremont Demonstrated “Reliable” NTV Operation, Enabled by
 - Novel circuit design techniques for logic, sequentials and memories
 - Variation aware design convergence strategies
- Next Steps:
 - Low overhead NTV circuits, ULV standard cell libraries
 - CAD Methodologies for Low Vcc & WDR designs : SSTA, Multipoint optimization
 - Device – Circuits – Architecture co-design for Near Threshold Computing

Q&A

Backup

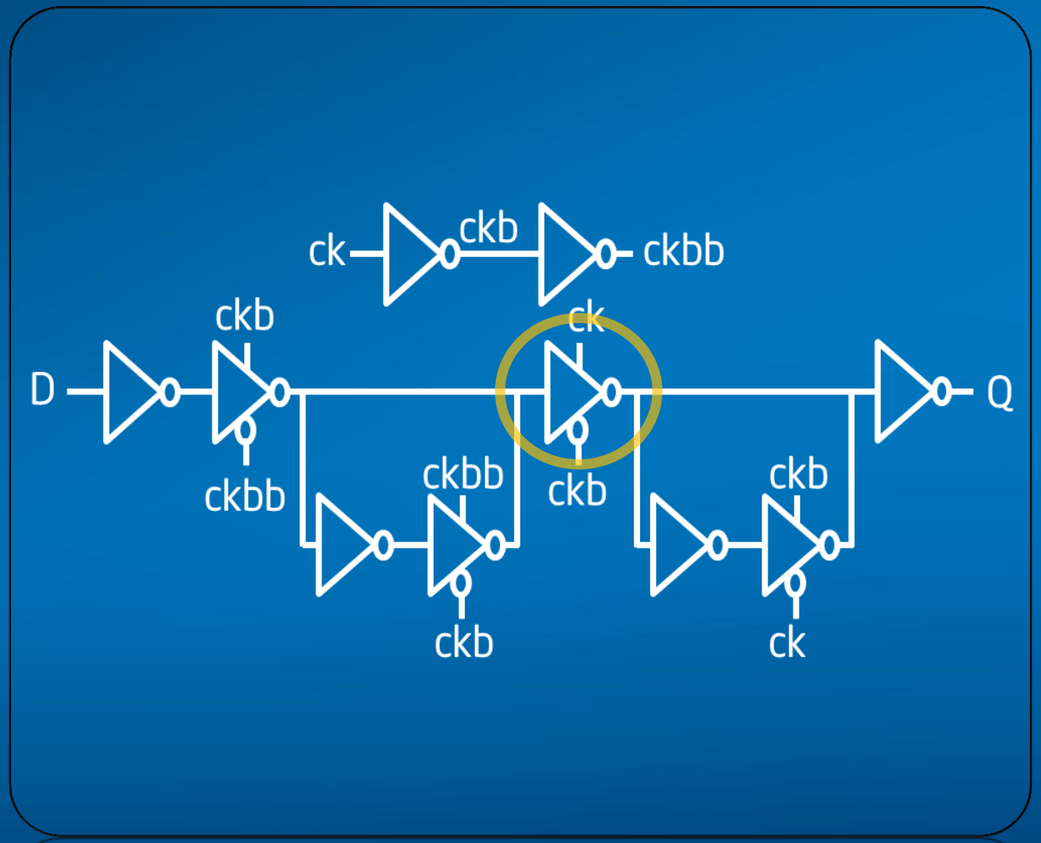
Near Threshold Voltage Logic



- Variation aware library pruning to ensure reliable NTV operation
- Limited transistor stacks to 3, No wide TG muxes, No contention circuits
- Pruned minimum sized and low drive strength cells: Minimum Z allowed is 2X process Z_{min}
- Only 40% of combinational standard cells in the library used in the design

Re-characterized constrained standard cell library at 0.5V NTV corner

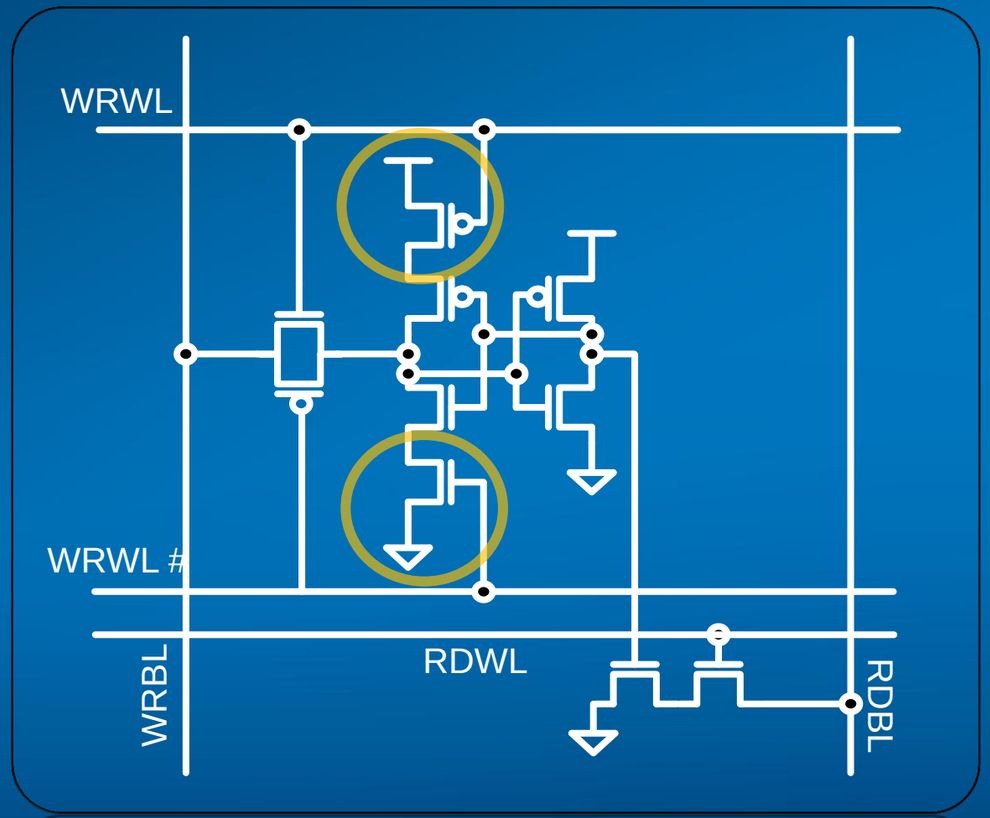
Near Threshold Voltage Sequentials



- Modified-C^2MOS topology with interruptible and upsized keepers
- Slave TG in conventional flip-flop replaced by clocked inverter, eliminating risk of write-back failures due to charge sharing
- Keepers upsized by 2X to achieve retention Vmin of 0.5V (RSSS, 5.5σ, -25^oC, 1.1MΩ Rg)
- Designed 13 custom sequential flavors and re-characterized at 0.5V

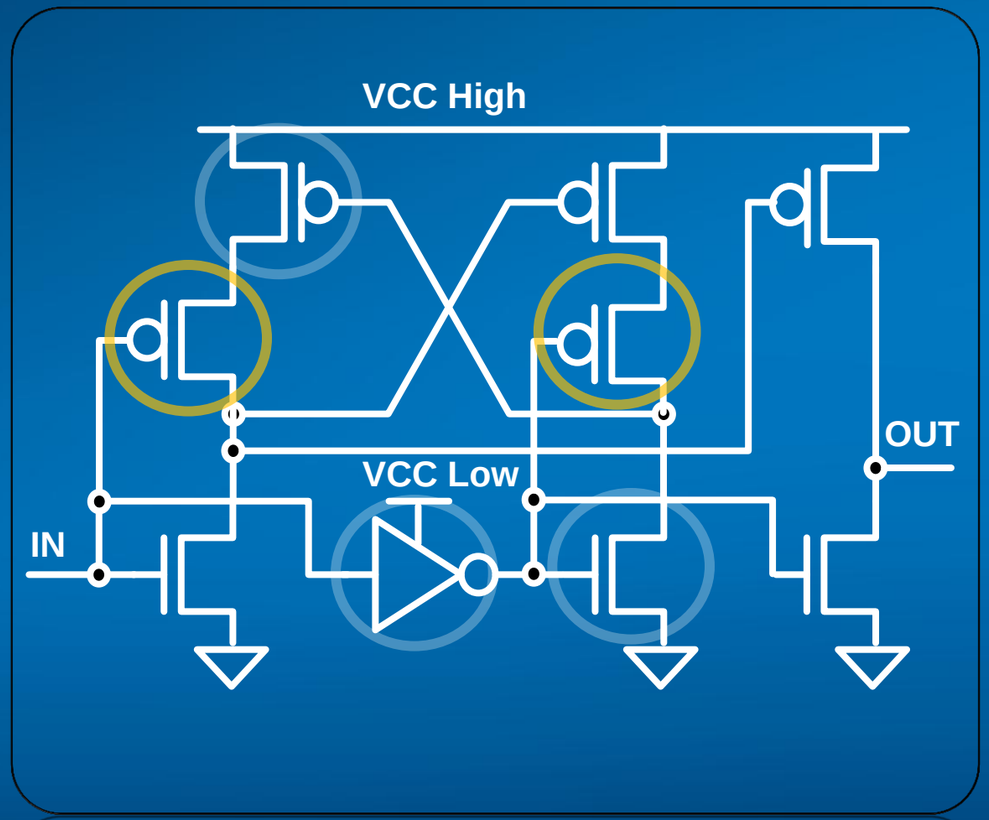
Robust, write-back free sequentials with 70mV retention Vmin improvement

Low Voltage RFs



- 10-T single ended transmission gate (SETG) latch topology
- Fully interruptible bit cell for contention free writes
- Retention limited cell, upsized (from 3-Track to 5-Track) to achieve retention Vmin of 550mV (RSSS, 5.9σ, -25ºC, 1.1MΩ Rg)
- Programmable keepers (3 vs. 4 stack) during read

Near Threshold Voltage Level Shifters



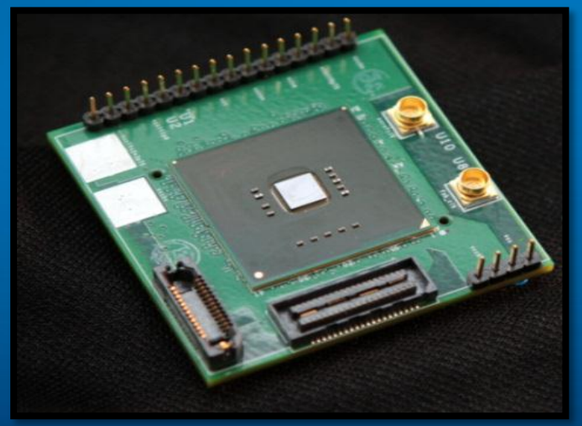
- Semi-interruptible split output level shifter topology
- Interruptible PMOS reduces contention, split output decouples delay path from latch stage
- Asymmetric critical path based sizing
- Two-stage level shifters between RLS & EBB with intermediate reference voltage
- Wide range level translation: sub-0.5V to 1.1V

60% performance improvement over single stage, symmetric level shifters

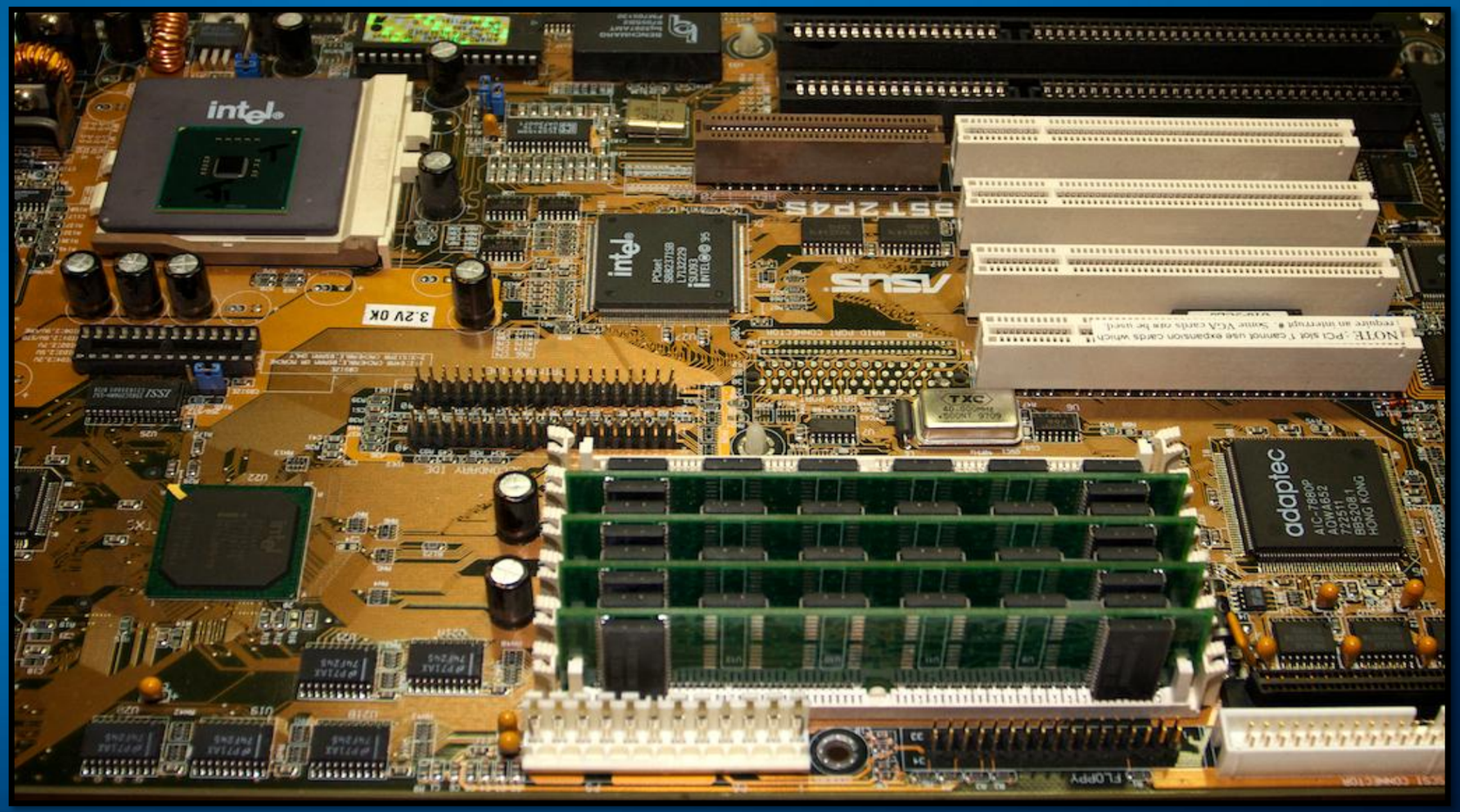
Package and Evaluation Board



951 Pin FCBGA Package



Custom Interposer



Legacy Pentium™ Socket-7 Motherboard

Legacy Benchmarks

Whetstone (inc. FPU):

- MWIPS MFLOPS VAXMIPS MWIPS-DP
- Pentium 100 66.2 16.8 97.8 66.2 1994
- Pentium 120 79.5 20.2 118 81.6 1995
- Pentium 133 88.3 22.4 130 90.8 1995

Dhrystone: (no FPU)

Dhry1 Dhry1 Dhry2 Dhry2
Opt NoOpt Opt NoOpt

- 80486 DX2 66 45.1 12.0 35.3 12.4
- Pentium 75 112 19.3 87.1 18.9
- Pentium 100 169 31.8 122 32.2
- Pentium 133 239 38.3 181 39.0
- Pentium 166 270 43.6 189 43.9

Linpack: (FPU heavy)
Opt No opt

- 80486 DX2 66 2.63 1.74
- Pentium 75 7.56 4.04
- Pentium 100 12.07 5.40
- Pentium 133 17.05 5.60
- Pentium 166 19.89 6.86

Reducing Transistor Variability For High Performance Low Power Chips

HOT Chips 24

Dr Robert Rogenmoser

Senior Vice President Product Development & Engineering

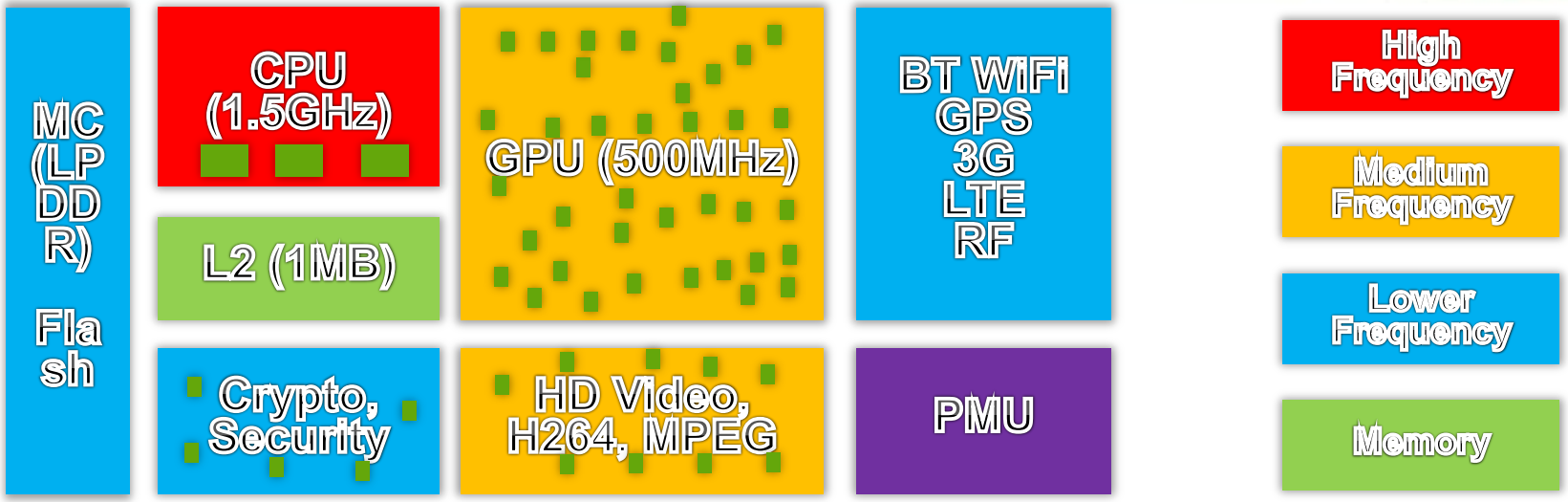
Copyright © 2012 SuVolta, Inc. All rights reserved.

Overview

- Transistor Variability Limits Chips
 - Impact on Mobile System on Chip (SOC)
 - Limited Low Power Design Techniques
 - Where does Variability come from?
- New Transistor Alternatives to Reduce Variability
 - Deeply Depleted Channel (DDC) technology
 - Silicon Impact
- Outlook
 - Taking advantage of Deeply Depleted Channel (DDC) in Mobile SOC

Copyright © 2012 SuVolta, Inc. All rights reserved.

What is needed in Mobile System on Chip?



- Multiple blocks with different performance requirements
 - Integrated on the same die
 - Different power modes – would like to run at different supplies
 - Multiple V_T transistors used to control leakage
 - Single chip solution requires analog integration
- Need co-design of architecture, circuits and transistor technology for best solution

Copyright © 2012 SuVolta, Inc. All rights reserved.

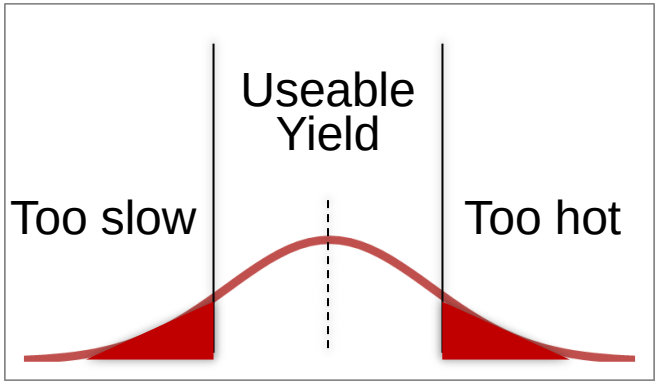
Variability Limits Design & Architecture

- Limited benefit using voltage scaling (DVFS)
 - Cannot overdrive much due to reliability and power restrictions
 - Dynamically lowering voltage limited to 100-200mV
 - Only lowering frequency leaves large leakage power
 - “Run to hold” beats DVFS despite overhead
- Finicky SRAM memories
 - High SRAM V_{MIN} leaves no room for memory voltage scaling
 - Many circuit tricks to improve V_{MIN} and noise margins
 - Design teams moved to dedicated power rail for SRAM
 - Works for CPU – difficult in GPU
 - Impacts power network integrity – more fluctuations
- Transistor variability limits chips

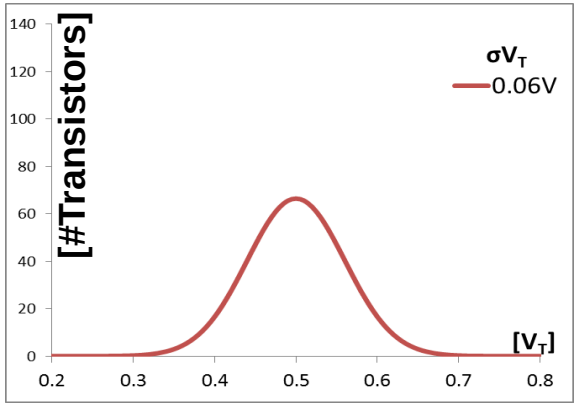
Copyright © 2012 SuVolta, Inc. All rights reserved.

Transistor Variation Source of Chip Variation

- Global/Systematic/Manufacturing Variation
 - Shifts all the transistors similarly
 - Longer/shorter transistor lengths
 - More (or less) implant energy and dose
 - ➔ Will result in speed/power distribution



- Local/Random Variation
 - Transistor next to each other vary widely
 - Small number of dopants in transistor channel
 - ➔ Random Dopant Fluctuation (RDF)
 - Apparent in threshold voltage mismatch (σV_T)
 - Impacts speed, leakage, SRAM & Analog



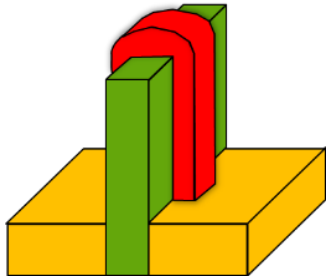
 - ➔ Industry solution: Remove RDF using **Undoped Channel**

- What is the right silicon roadmap going forward?

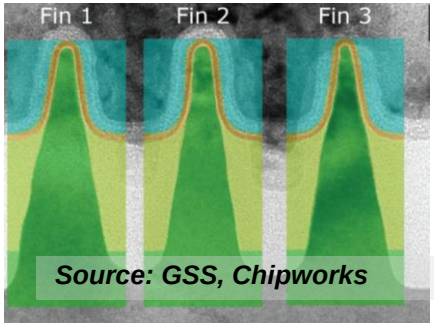
Copyright © 2012 SuVolta, Inc. All rights reserved.

Transistor Alternatives

- FinFET or TriGate
 - Promises high drive current
 - Manufacturing, cost, and IP challenge
 - Doped channel to enable multi V_T

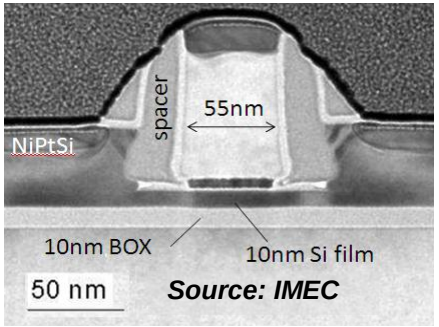


Textbook FinFET

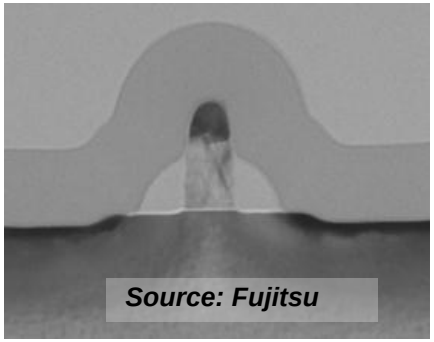


Intel TriGate

- FDSOI
 - Showing off undoped channel benefits
 - Good body effect, but lack of multi V_T capability
 - Restricted supply chain



- DDC – Deeply Depleted Channel transistor
 - Straight forward insertion into Bulk Planar CMOS
 - Undoped channel to reduce random variability
 - Good body effect and multi V_T transistors

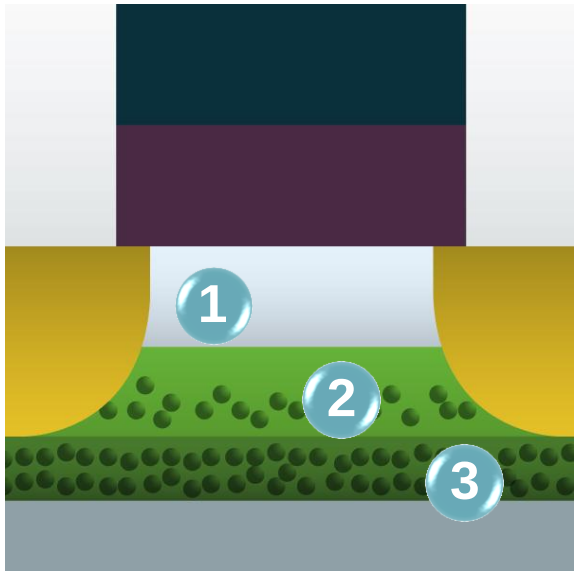


Copyright © 2012 SuVolta, Inc. All rights reserved.

Deeply Depleted Channel™ (DDC) Transistor

1 Undoped or very lightly doped region

- Significantly reduced transistor random variability σV_T
 - ➔ Lower leakage
 - ➔ Better SRAM (I_{READ}, lower V_{min} & V_{ret})
 - ➔ Tighter corners
 - ➔ Smaller area analog design
- Higher channel mobility (increased I_{eff}, lower DIBL)
 - ➔ Higher speed, improved voltage scaling



**Example implementation*

2 V_T setting offset region

- Enables multiple threshold voltages

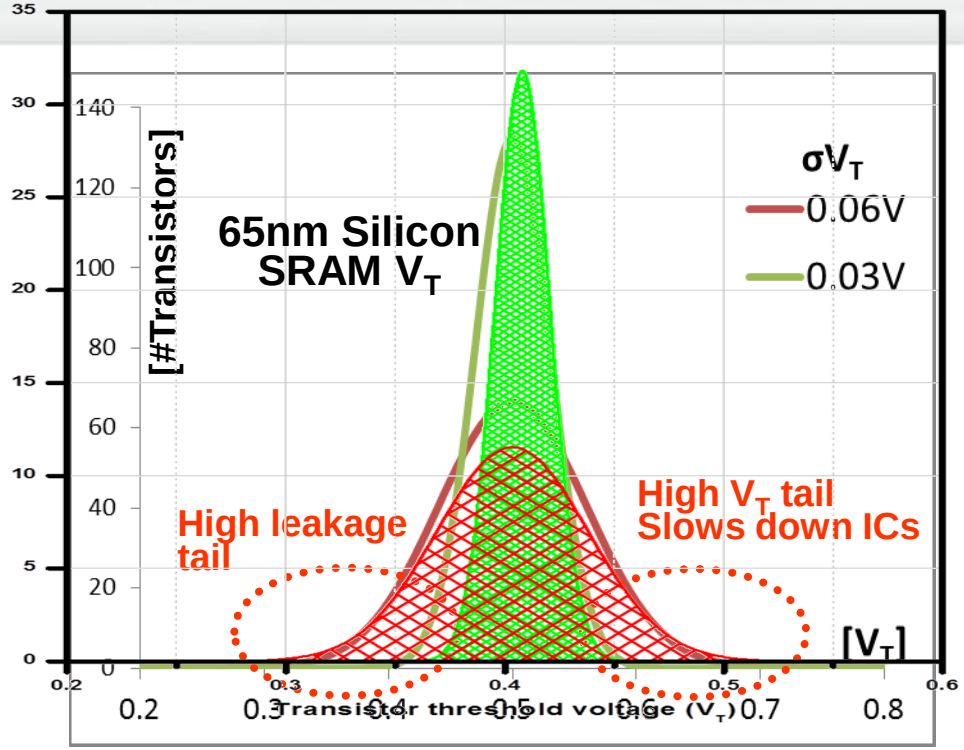
3 Screening region

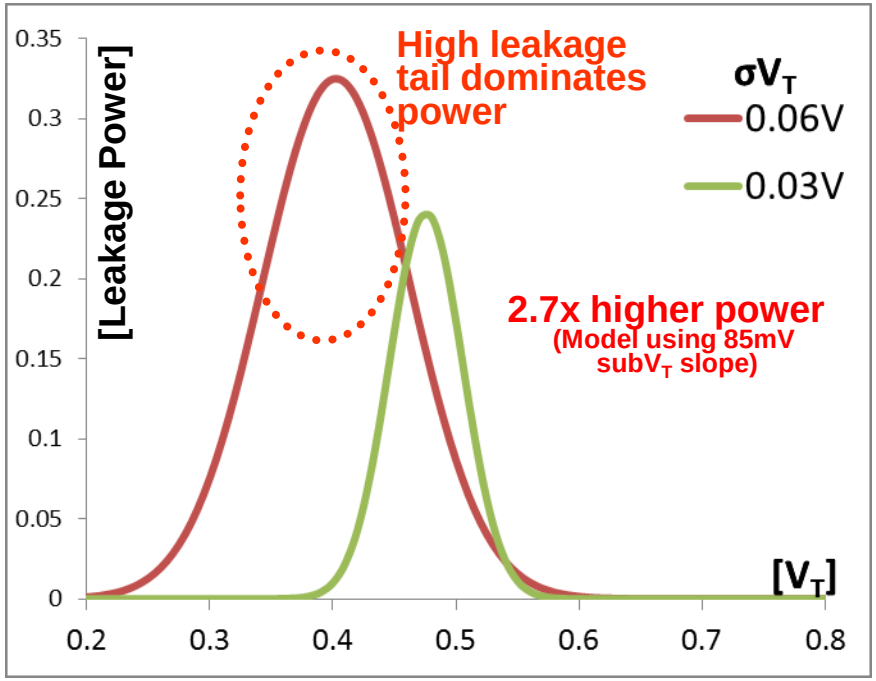
- Strong body coefficient
 - ➔ Bias bodies to tighten manufacturing distribution
 - ➔ Body biasing to compensate for temperature and aging

Benefits similar to FinFET in planar bulk CMOS

Copyright © 2012 SuVolta, Inc. All rights reserved.

Lower Transistor Variability Reduces Leakage

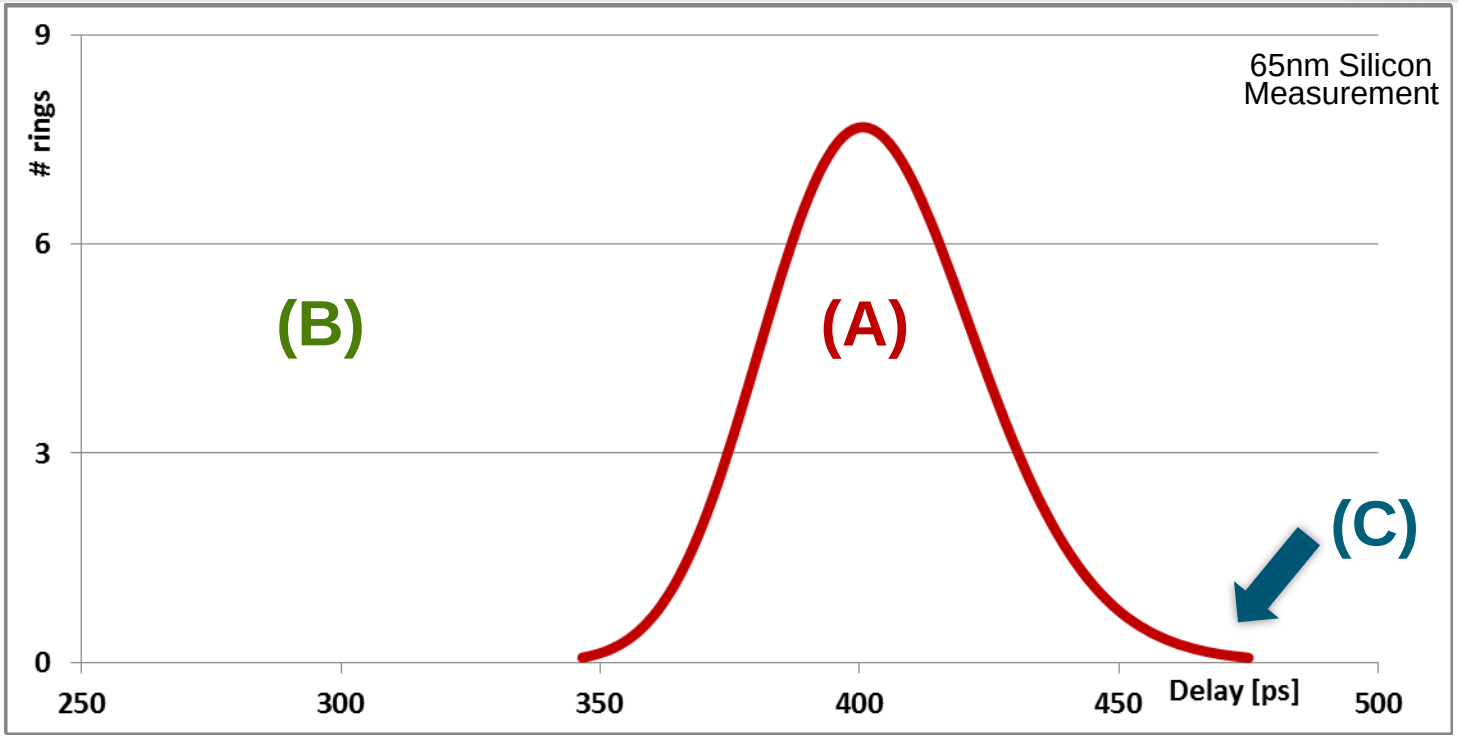




- Transistor variability is reflected in threshold voltage (V_T) distribution
- Leakage current is exponentially dependent on V_T
- Lower V_T variability (σV_T) reduces number of leaky low V_T devices
- Power dissipation is dominated by low V_T edge of distribution
- Smaller σV_T → Less leakage power for digital and memory/SRAM

Copyright © 2012 SuVolta, Inc. All rights reserved.
SuVolta®

Lower Transistor Variability Improves Speed

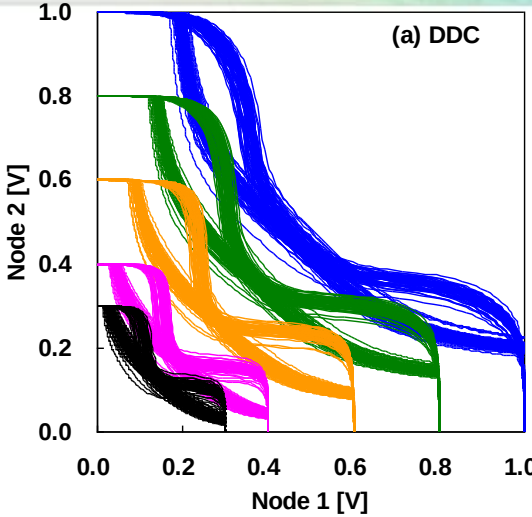


- Nominal (TT) ring oscillator speed expected to be 400ps **(A)**
 - Equivalent to having many similar critical paths in a chip
 - V_T variation will randomly affect paths within the same die limiting speed to 470ps
- Undoped channel reduces variability and increases mobility **(B)**
 - 25% faster mean, 30% faster tail due to tighter distribution
- To match performance lower V_{DD} until tails have same speed **(C)**
 - Large impact on power due square dependence $P=CV^2f + IV$

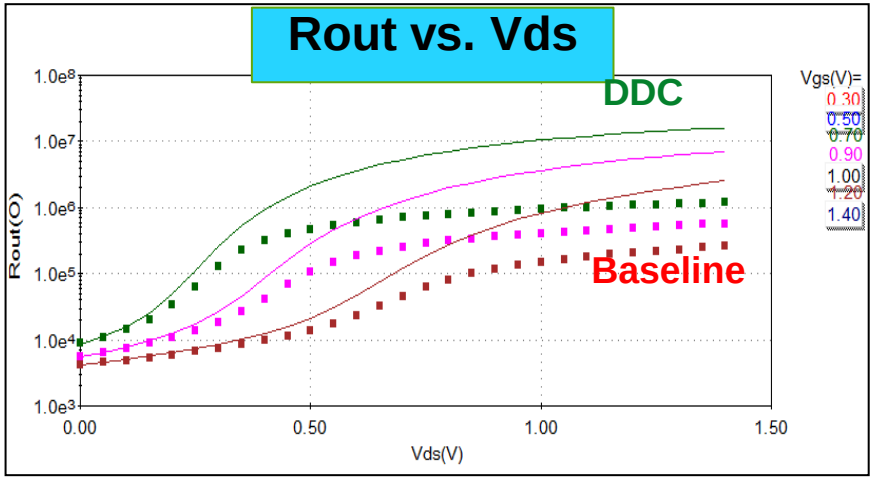
Copyright © 2012 SuVolta, Inc. All rights reserved.

Lower Variability Improves Transistor Matching

- SRAM memories built using 6-T SRAM cell
 - Smallest transistors on every chip, worst V_T mismatch
 - Higher V_{DD} is required to avoid failures
 - Demonstrated SRAM to V_{min} of 0.425V
- In analog circuits, matching is key
 - Large transistors used to improve relative variability in current mirrors, differential pairs, etc.
 - Better transistor matching allows for
 - Area savings
 - Higher performance
 - Lower power
 - Undoped channel improves R_{OUT} ➔ higher gain

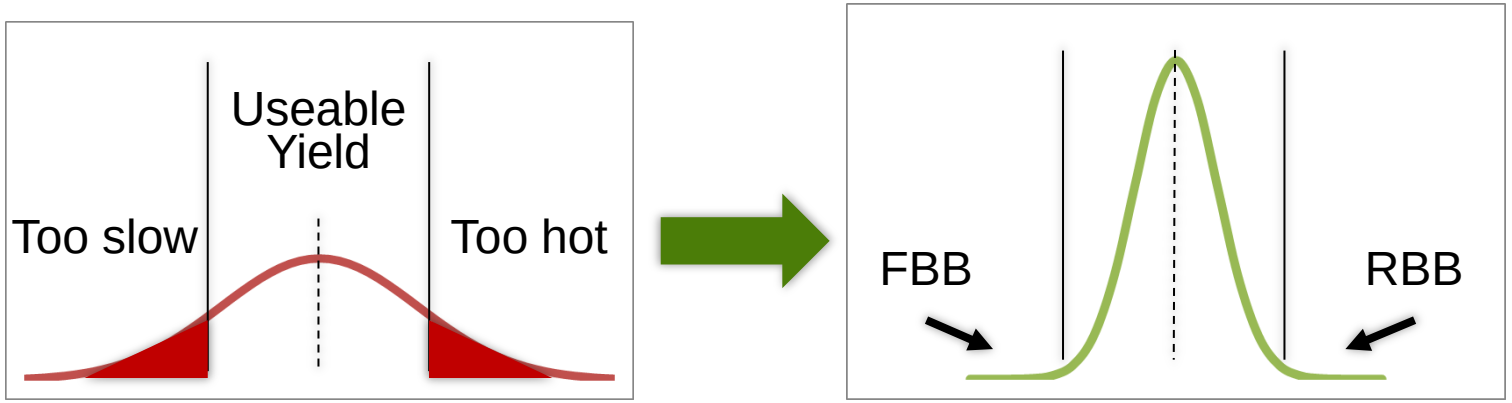


65nm Silicon Measurement



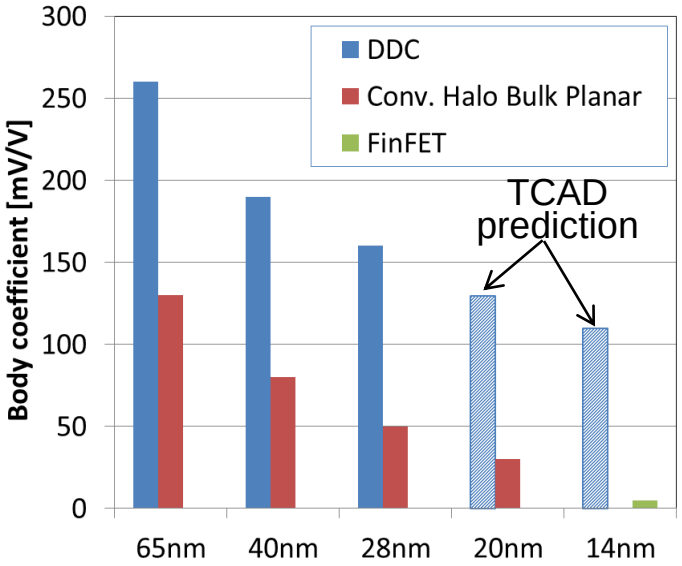
Copyright © 2012 SuVolta, Inc. All rights reserved.
SuVolta®

Better Chips with Body Biasing



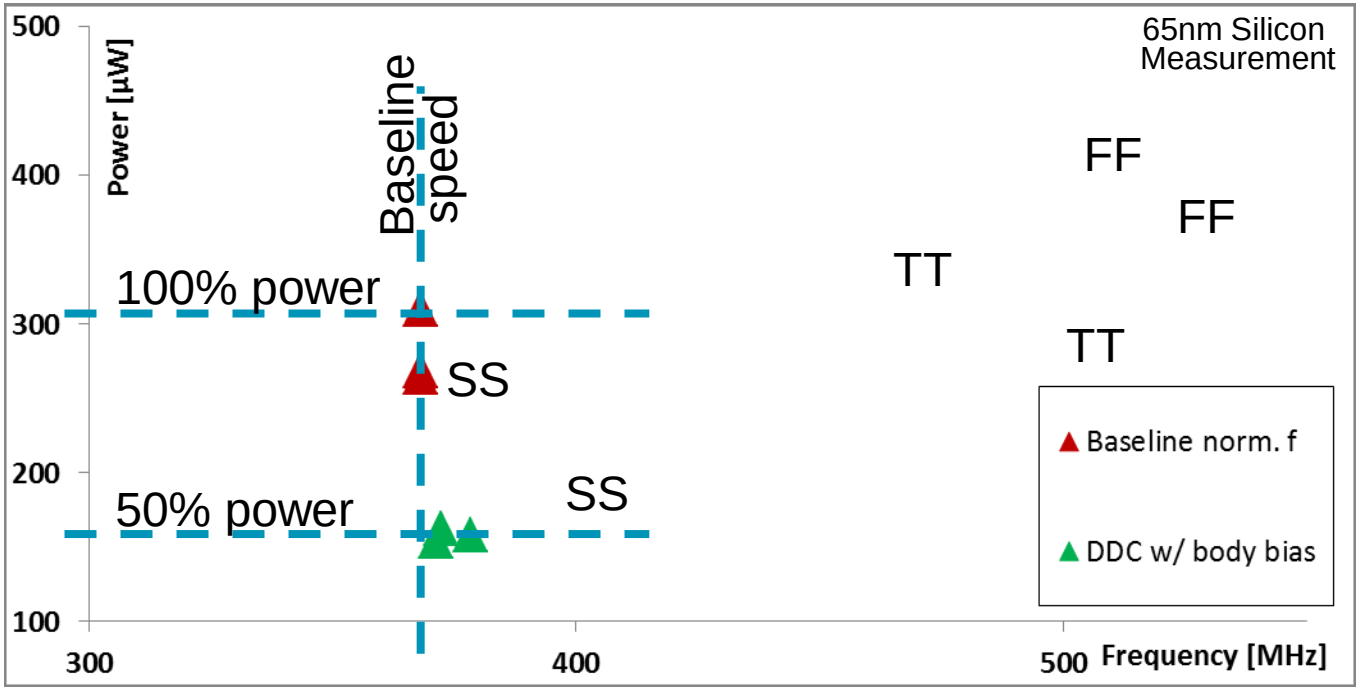
- Body Bias to fix systematic variation
 - Speed-up (forward bias - FBB) slow parts
 - Cool down (reverse bias - RBB) hot parts

 ➔ Increase manufacturing yield
- Body bias enables multiple modes of operation
 - Active → minimize power at every performance
 - Standby → leakage reduction, power gating
- DDC provides 2-4x larger body factor



Copyright © 2012 SuVolta, Inc. All rights reserved.
SuVolta®

Half the Power at Matched Performance

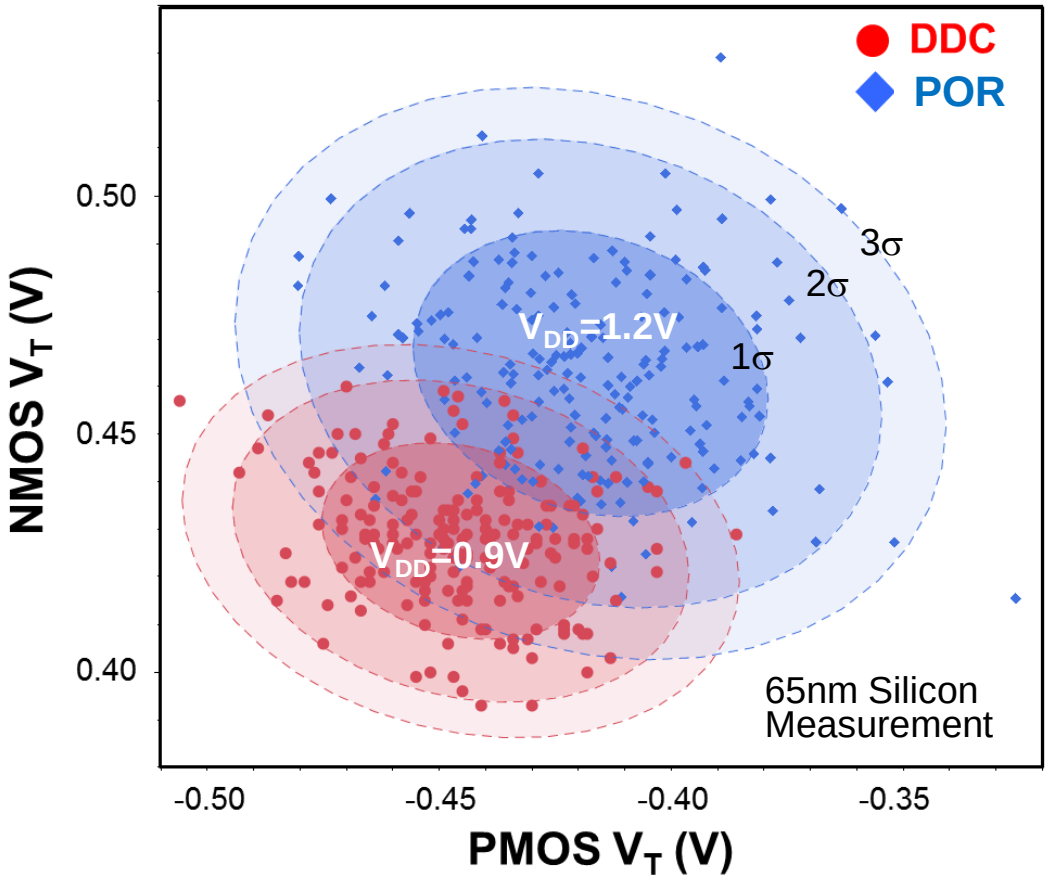


- Inverter ring-oscillators (RO) fabricated at process corners
 - Baseline @ 1.2V V_{DD} and DDC @ 0.9V V_{DD}
- For each corner, DDC RO is faster and lower power
- Using strong body coefficient to pull in corners
 - Half the power (50% less power) while matching speed

Copyright © 2012 SuVolta, Inc. All rights reserved.

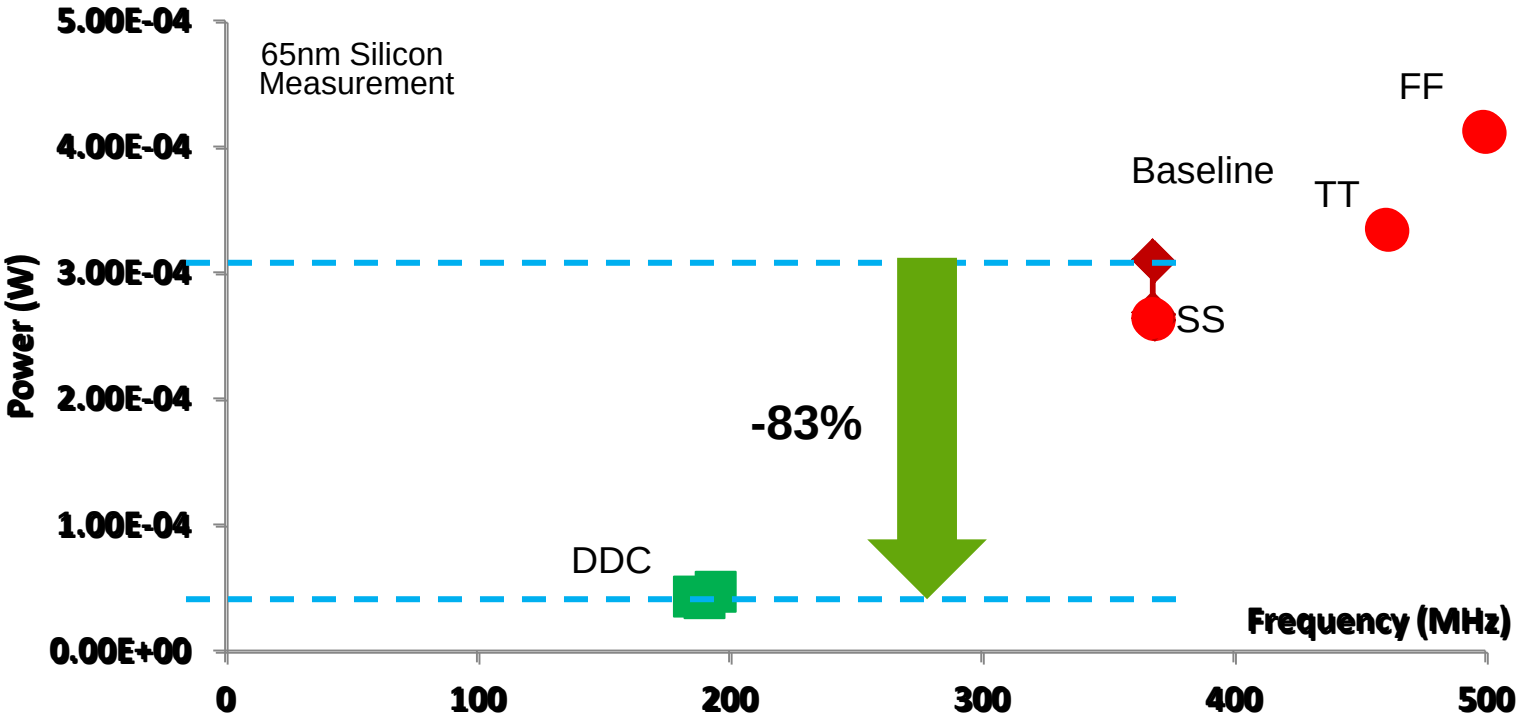
Tighter Manufacturing Corners w/ DDC

- Better process control leads to tighter corners
 - Manufacturing flow further reduces layout effects
- 1 sigma tighter wafer to wafer and within wafer variation for DDC
- Less overdesign as max paths and min (hold) paths are closer
- Faster design closure
 ➔ earlier tapeout
 ➔ shorter TTM



Copyright © 2012 SuVolta, Inc. All rights reserved.
SuVolta®

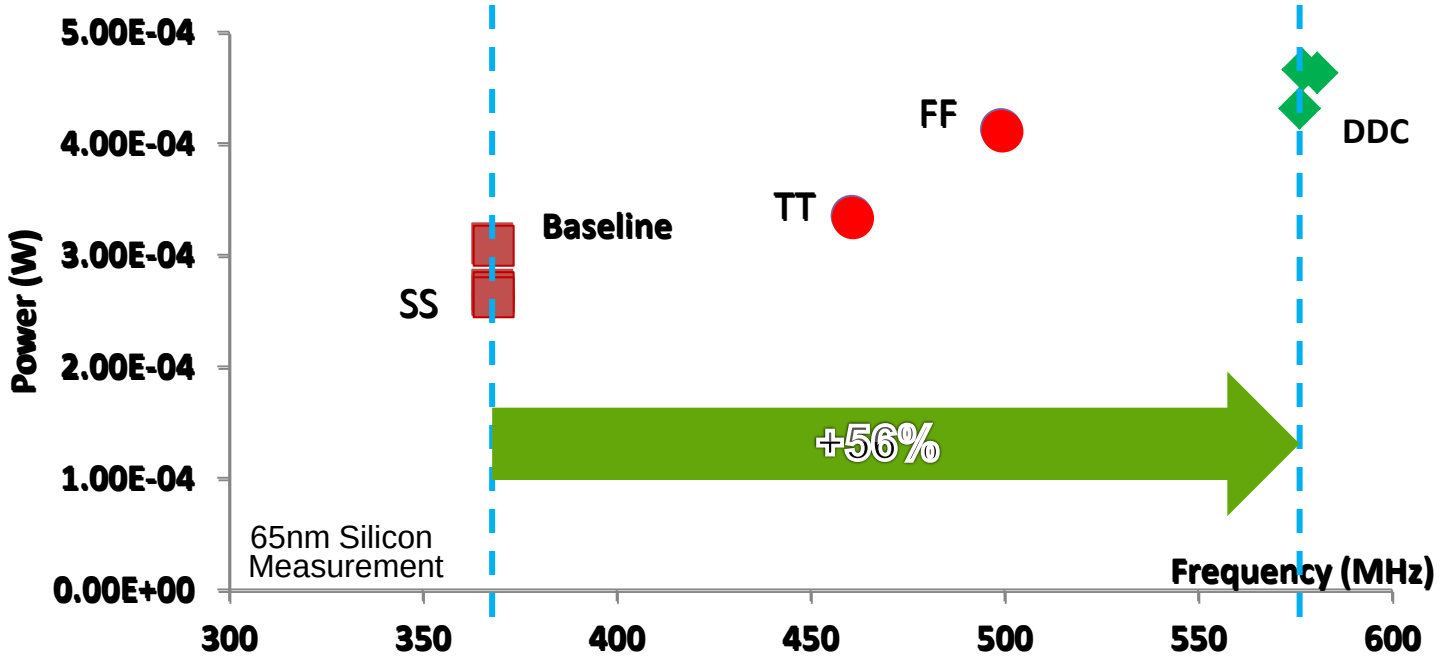
Voltage Scaling to 0.6V V_{DD}



- Achieve half the speed at 1/6 the power @0.6V V_{DD}
- Use body bias to compensate for temperature and aging
 - Critical for low V_{DD} operation
 - Enable workable design window – avoid overdesign

Copyright © 2012 SuVolta, Inc. All rights reserved.

This is HotChips – Go Faster!

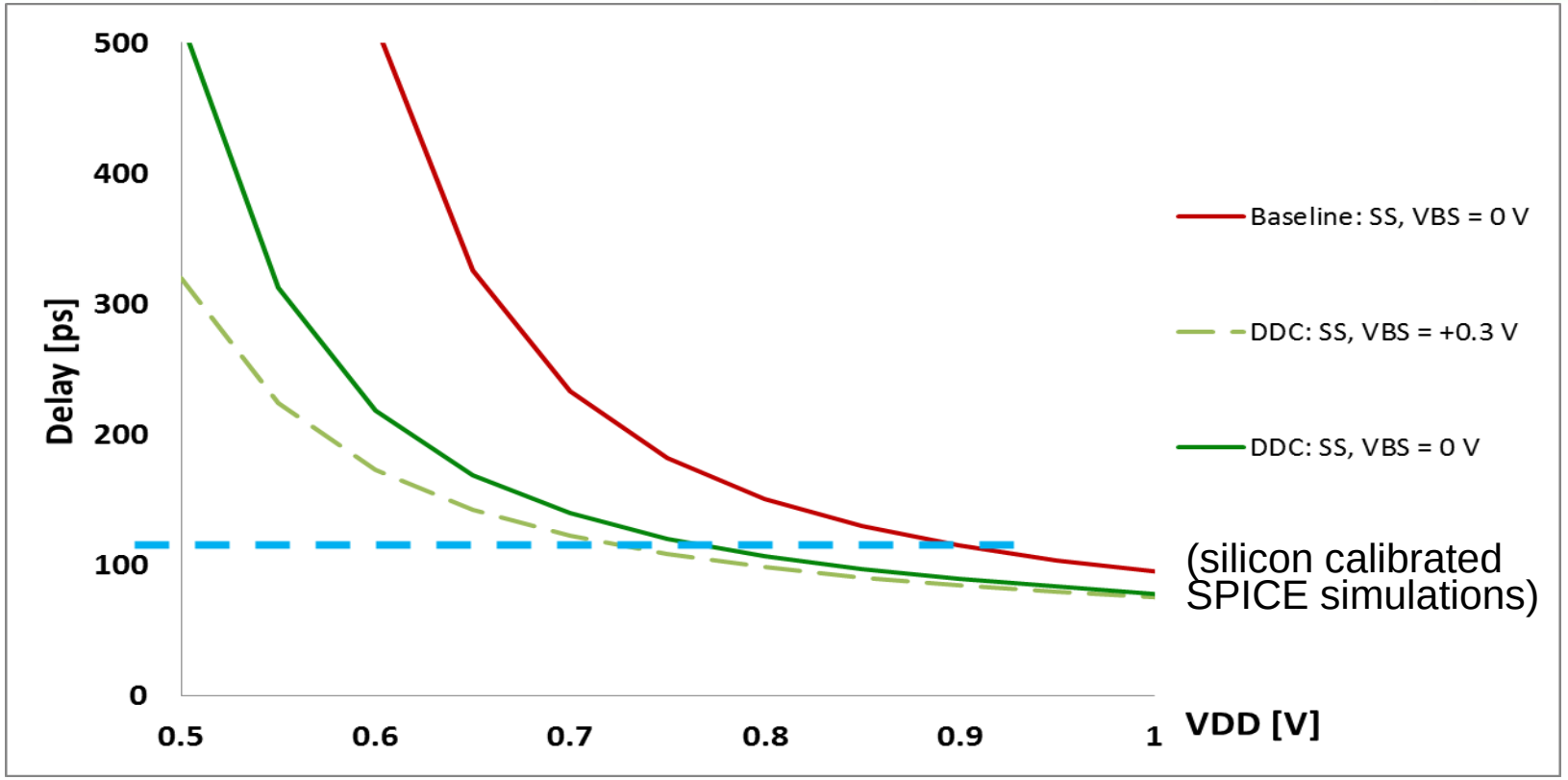


- Turbo Mode: DDC achieves over 50% speedup @ 1.2V V_{DD}
 - All corners for DDC run at 580MHz vs 370MHz for baseline

DVFS	Baseline	DDC			
V_{DD}	1.2V	0.6V	0.9V	1.05V	1.2V
Speed	**1**	0.5	**1**	1.28	1.56
Power	**1**	0.17	0.52	**1**	1.51

Copyright © 2012 SuVolta, Inc. All rights reserved.

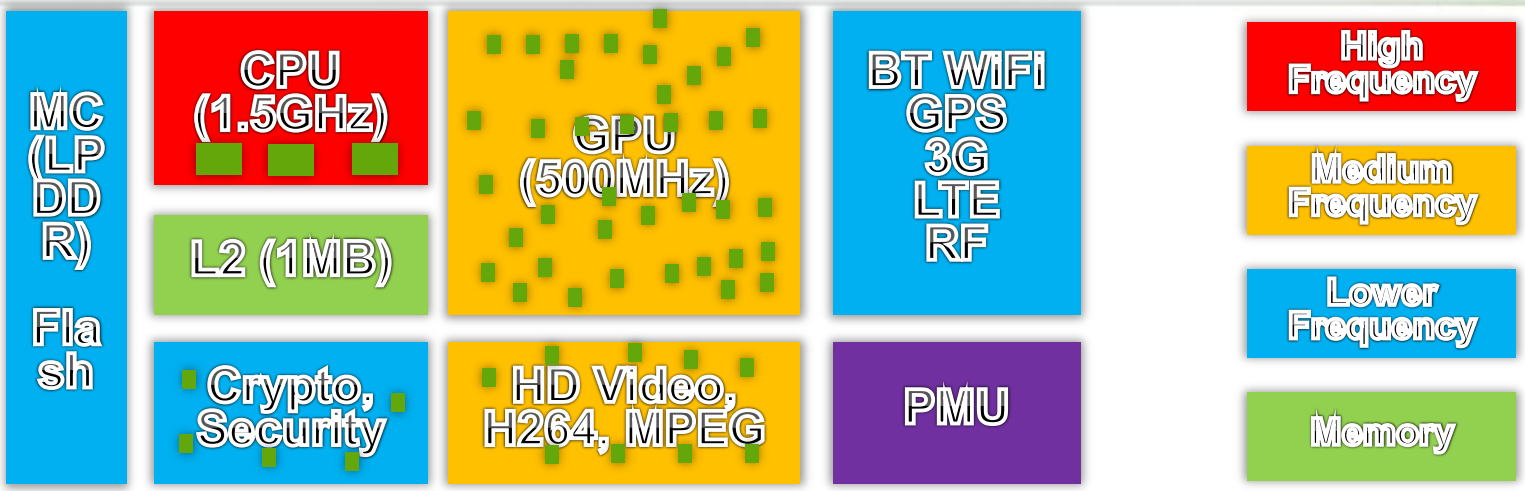
28nm and Beyond



- Same performance at 0.75V V_{DD} as baseline at 0.9V V_{DD}
 - 30% lower power
 - Alternatively 25% faster at same voltage
- Even better when using body bias to pull in corners

Copyright © 2012 SuVolta, Inc. All rights reserved.

Applying DDC to Lower Variability in Mobile SOC



- CPU: Single thread performance critical
 - Push frequency by temporarily raising voltage in turbo mode
 - DVFS with body biasing becomes DVBFS
- GPU: High number of cores using small transistors
 - Less overdesign due to lower delay variability
 - Increase parallelism, lower voltage, body bias dynamically for more pixels/Watt
- Lower frequency blocks
 - In addition to high V_T transistors also run at lower voltage and optimal body bias
- Whole chip: Use body bias to adjust for manufacturing variation
 - Take advantage of improved memory and analog performance
 - Lowering variability while compatible with existing bulk planar silicon IP

Copyright © 2012 SuVolta, Inc. All rights reserved.

Conclusions

- Variability limits chips
 - DDC reduces random variability through its undoped channel
 - DDC's strong body factor can be used to fix systematic variation and compensate for temperature variation
- DDC provides performance kicker from 90nm to 20nm
 - Straight forward integration into existing nodes
 - Compatible with existing bulk planar CMOS silicon IP
 - Use existing CAD flow
- DDC brings back low power tools
 - Large range DVFS
 - Body biasing
 - Low voltage operation
- Taking advantage of reduced variability DDC in design and architecture will lead to next level in mobile SOC

Copyright © 2012 SuVolta, Inc. All rights reserved.



Copyright © 2011 SuVolta, Inc. All rights reserved.



High performance and efficient single-chip small cell base station SoC

Kin-Yip Liu
Cavium, Inc.
kliu@cavium.com

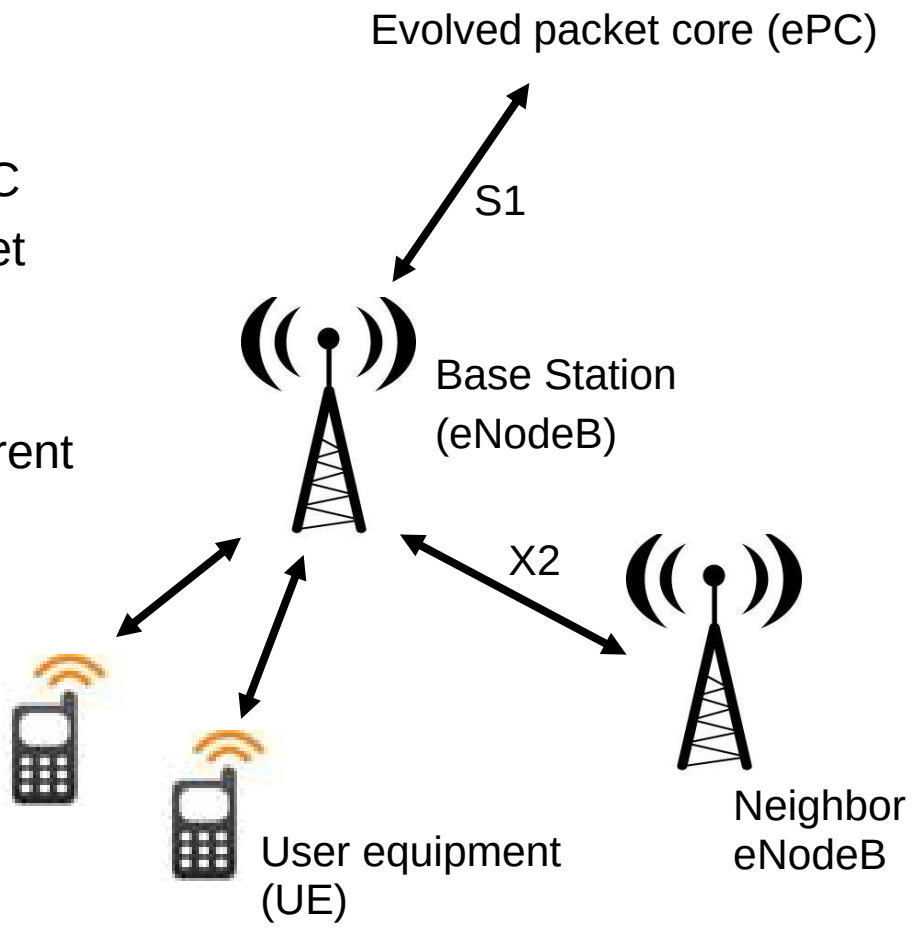
Hot Chips 24, August 2012

Presentation Overview

- Base station processing overview
- Why small cells and heterogeneous Radio Access Network (RAN)
- Small cell design based on OCTEON Fusion
- OCTEON Fusion CNF71XX architecture
- CNF71XX design
- Software models
- Summary

LTE Wireless Network Overview

- LTE equipment:
 - Base Stations – eNodeB
 - User equipment (UE), e.g. cell phone, dongle for notebook PC
 - Core network – Evolved Packet Core (ePC)
- An eNode interfaces with:
 - ePC (multiple nodes with different functions)
 - Control, signaling
 - To voice & data networks
 - UE's
 - Neighbor eNodeB's
 - Communicate load and interference info
 - Handover UE's



LTE Protocols & Processing

- eNodeB relays information between UE and ePC
- eNodeB and UE communication protocol:

Protocol layers	Processing functions
RRC (layer 3)	Set up and maintain radio bearers. Manage radio resources. Control functions. Handover decisions
PDCP (layer 2)	En/decrypt over-the-air traffic, Header de/compression
RLC (layer 2)	Segment and reassemble traffic. Ensure in-order traffic delivery. Re-transmit as needed
MAC (layer 2)	Schedule use of over-the-air resources. Select PHY configuration for transfers. Collect stats & report to RRC
PHY (layer 1)	Physical layer: OFDM for downlink. SC-FDMA for uplink

- eNodeB and ePC communication protocol:
 - IP network, IPSec protected, GTP tunnels of user data in UDP/IP, SCTP for control traffic

Classes of Base Stations

CAVIUM

Small Cells

	Home Femto	Enterprise Femto	Pico	Micro	Macro
Cell Radius	50m	75m	250 - 400m	2 - 20km	20km
No. of users	8	32	128	1200	3600
Peak data rate	50Mbps DL 25Mbps UL	100Mbps DL 50Mbps UL	150Mbps DL 75Mbps UL	300Mbps DL 150Mbps UL	900Mbps DL 450Mbps UL
User Mobility	4 km/hr	4 km/hr	50 km/hr	350 km/hr	350 km/hr
Locations	Home	Office, school, apartment buildings, malls	Urban hotspots, rural areas	Urban, rural areas	Metro, traditional approach

DL – Downlink. Traffic going from network to user
UL – Uplink. Traffic going from user to network

Additional Small Cell Requirements

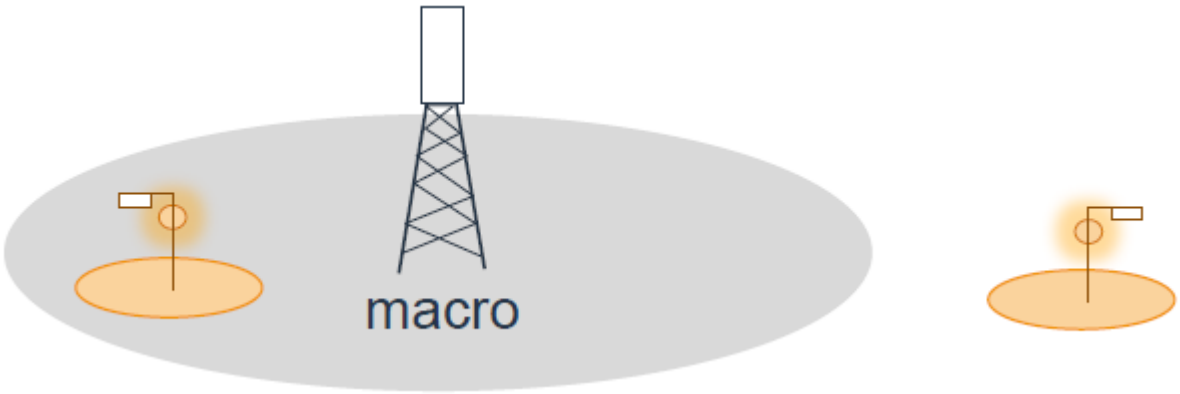
CAVIUM

- WiFi option
 - Single platform for Small Cell + Access Point
 - SoC must provide performance headroom for both functions
- Power-over-Ethernet
 - Simplify system deployment, but limited system power supply
 - SoC must consume very low power
- Time synchronization
 - Mandatory for LTE base stations. IP backhaul, no TDM interface
 - GPS option. May not work well in-door
 - Software solutions: IEEE 1588 v2, NTP. In-door OK, cost effective
- Security
 - Authenticated and encrypted software for secure boot

Why deploy small cells?

CAVIUM

…….for **Hot spots** and **Not spots**



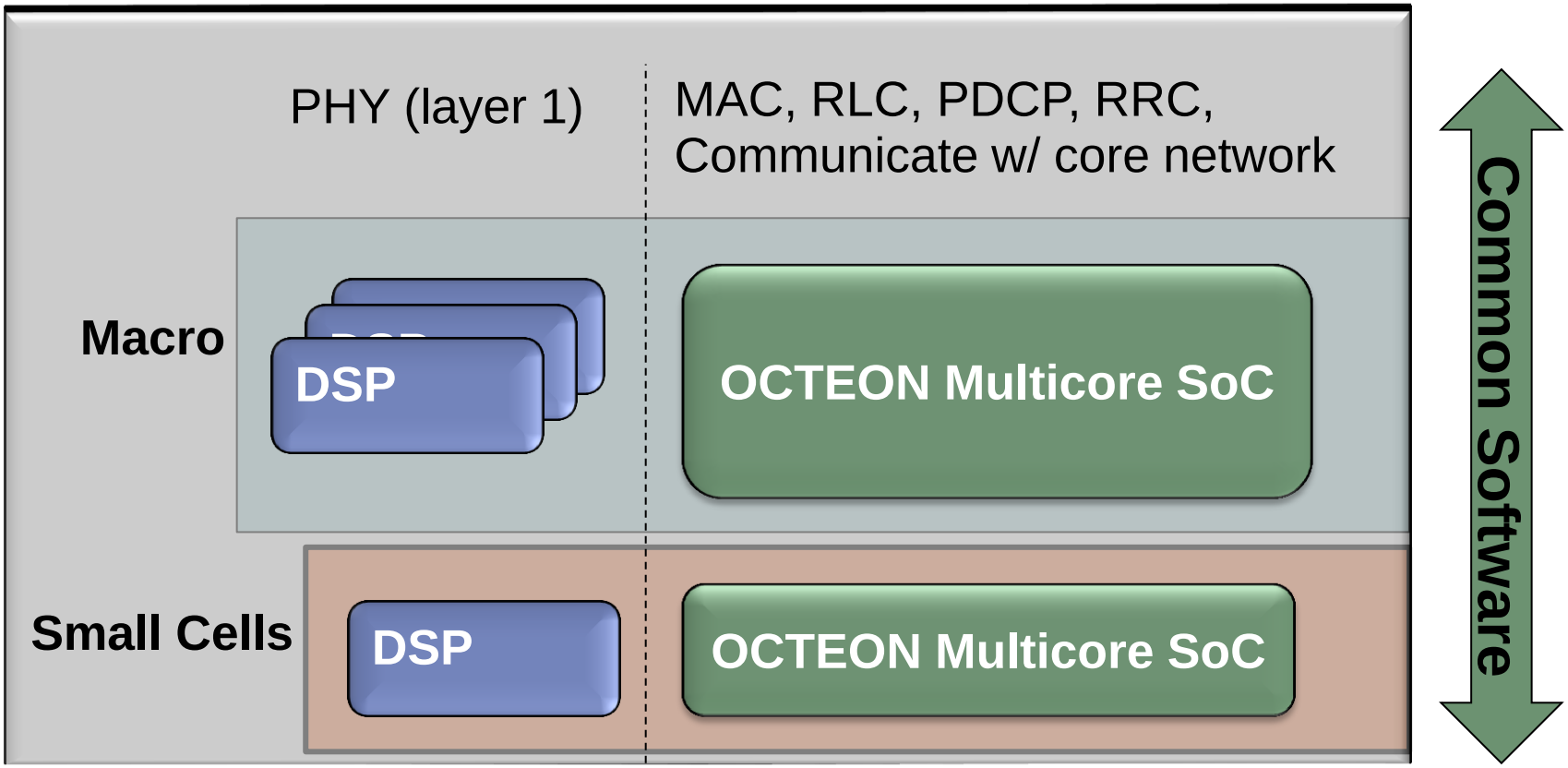
Easing congestion
within macro coverage

New coverage
in addition to macro

Small Cells essential for LTE coverage, capacity, and throughput

Current Generation Base Stations

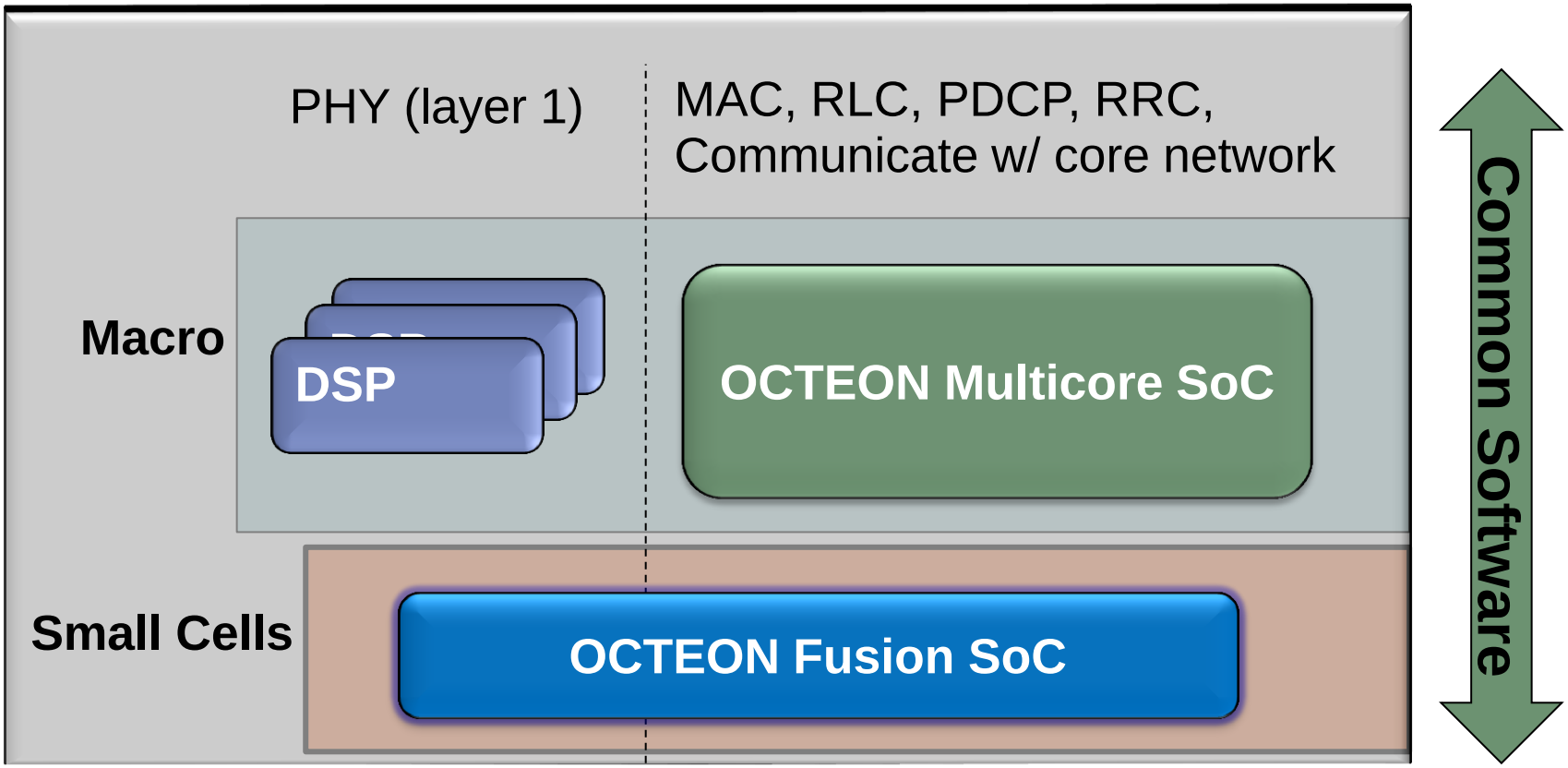
CAVIUM



Single-chip Multicore SoC for Layer 2 and above processing. Common software from Small to Macro cells

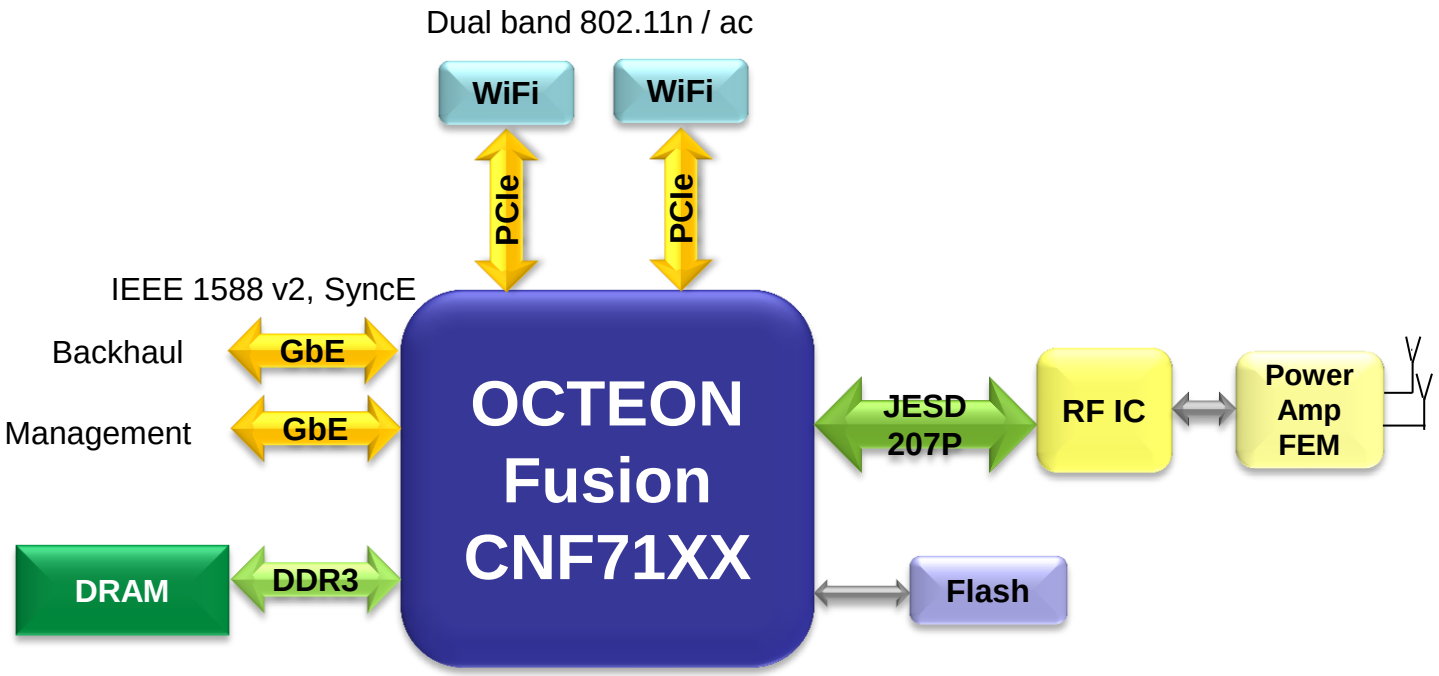
Next Generation Base Stations

CAVIUM



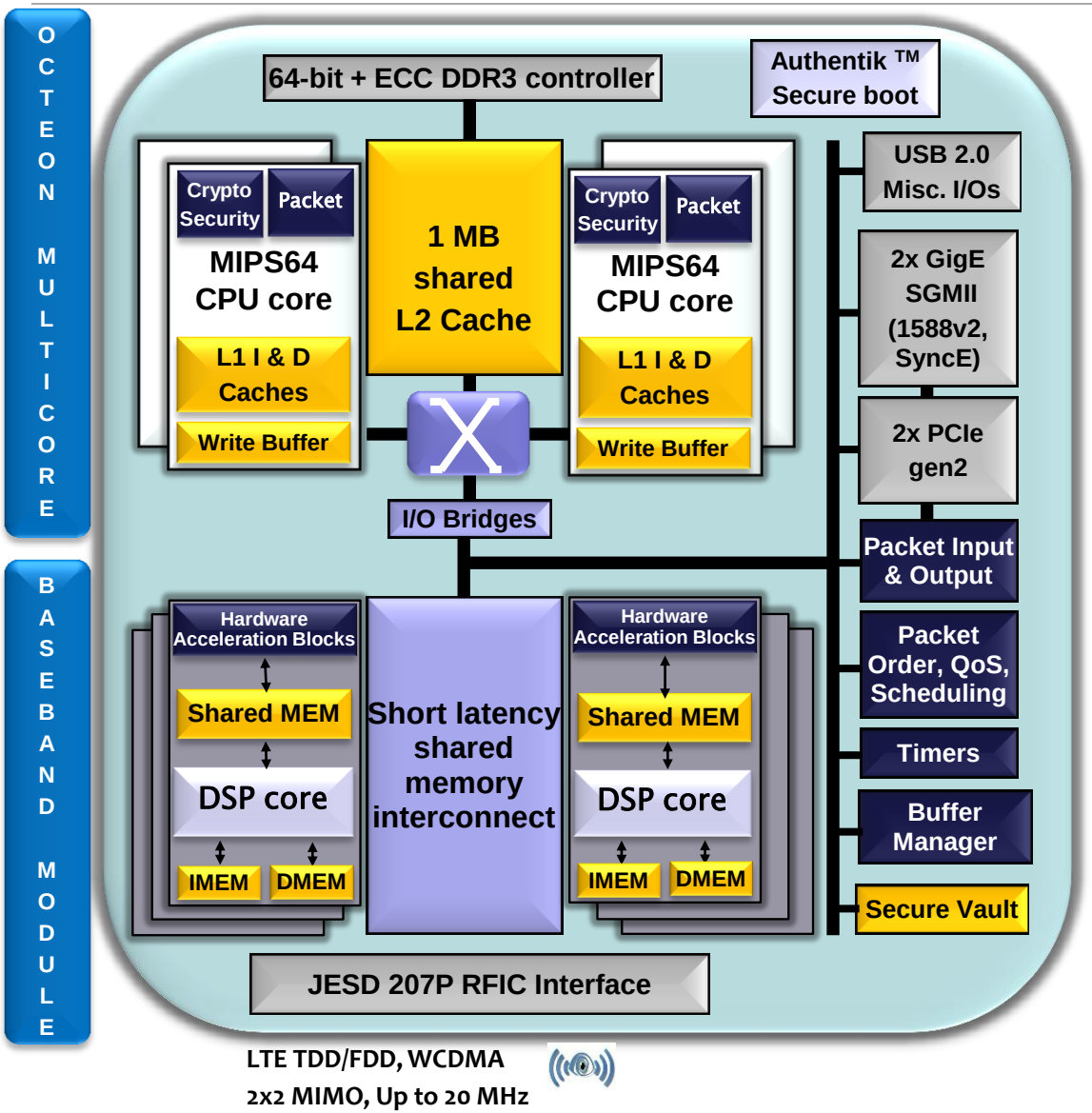
Single-chip Multicore + baseband module SoC for Small Cells. Common software from Small to Macro cells

OCTEON Fusion based Small cell



Small Cell Base Station + Access Point

OCTEON Fusion CNF71XX
Small cell BaseStation-on-a-chip Family



- **High Performance LTE / 3G Small Cell SoC Processors:**
 - 4 MIPS64 cores up to 1.5 GHz
 - 6 DSP cores up to 500MHz
 - Many HW Accelerators for Packet Processing, LTE/3G, and Security
 - IEEE 1588 v2, SyncE
 - Authentik secure boot
- **Highly Scalable**
 - Spanning 32 to 200+ Users
 - 3G and LTE FDD & TDD
 - Up to LTE 20MHz 150 Mbps Uplink (UL) + 150Mbps Downlink (DL)
- **Headroom for Unique Carrier Class Features**
 - Multi-User MIMO
 - Self Optimizing Networks
 - Interference Cancellation
 - Advanced Receivers

Design Philosophy

CAVIUM

High Performance and Power Efficient	• Power and area efficient CPU and DSP cores • Scale performance with more cores • Not depend on very high frequency or core complexity
Short Latencies Deterministic Performance	• Shortest cache and memory latencies. Optimize for determinism • Flexible prefetch, cache hints, options to cache packet headers only • L2 way partition feature avoids cache pollution
Optimized ISA Ease of programming	• MIPS64 r3 instruction set + >80 OCTEON instructions • Full C programming. Standard OS and development tools
Comprehensive Hardware Acceleration	• TCP/IP, complete packet receive and transmit offload, packet ordering, QoS, work scheduling, buffer de/allocation, IPSec, wireless crypto algorithms, timers, wireless baseband functions • Crypto coprocessor in each core. Best latency & determinism
Software Compatible Roadmap	• Software compatible from 1-48 cores and across generations • Single SDK to develop software for all OCTEONs • Software for macro base stations directly reusable for Small Cells

Baseband Module

CAVIUM

Baseband module processing flows

- Wireless UL and DL processing differ. Partition the DSP cores and assign relevant hardware accelerators for UL Vs. DL processing
- Modular design with flexible partitioning simplifies software design

6x DSP cores optimized for wireless baseband processing

- 3-way VLIW, with 16x MAC or 4x complex MAC vector processing per cycle
- Optimizing instructions for wireless baseband processing
- Dual 128-bit load/store paths transfer up to two vector operands each cycle

Hardware accelerators (HABs)

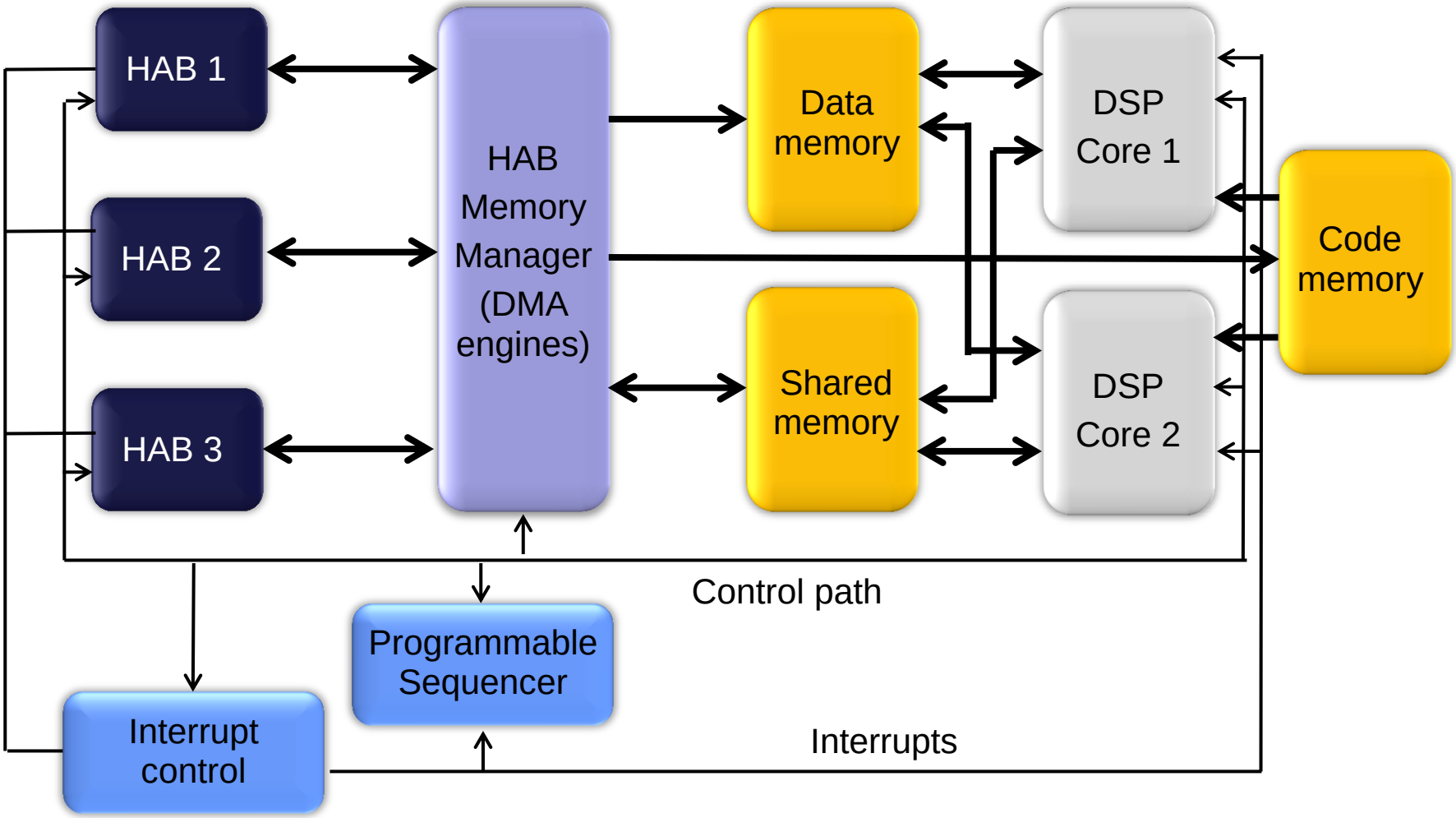
- Comprehensive set of LTE and 3G, UL and DL relevant accelerators
- Automate offload to accelerators with DMA engines and Sequencer

Shared memory interconnect

- DSPs and HABs can access any memory structure in entire baseband module

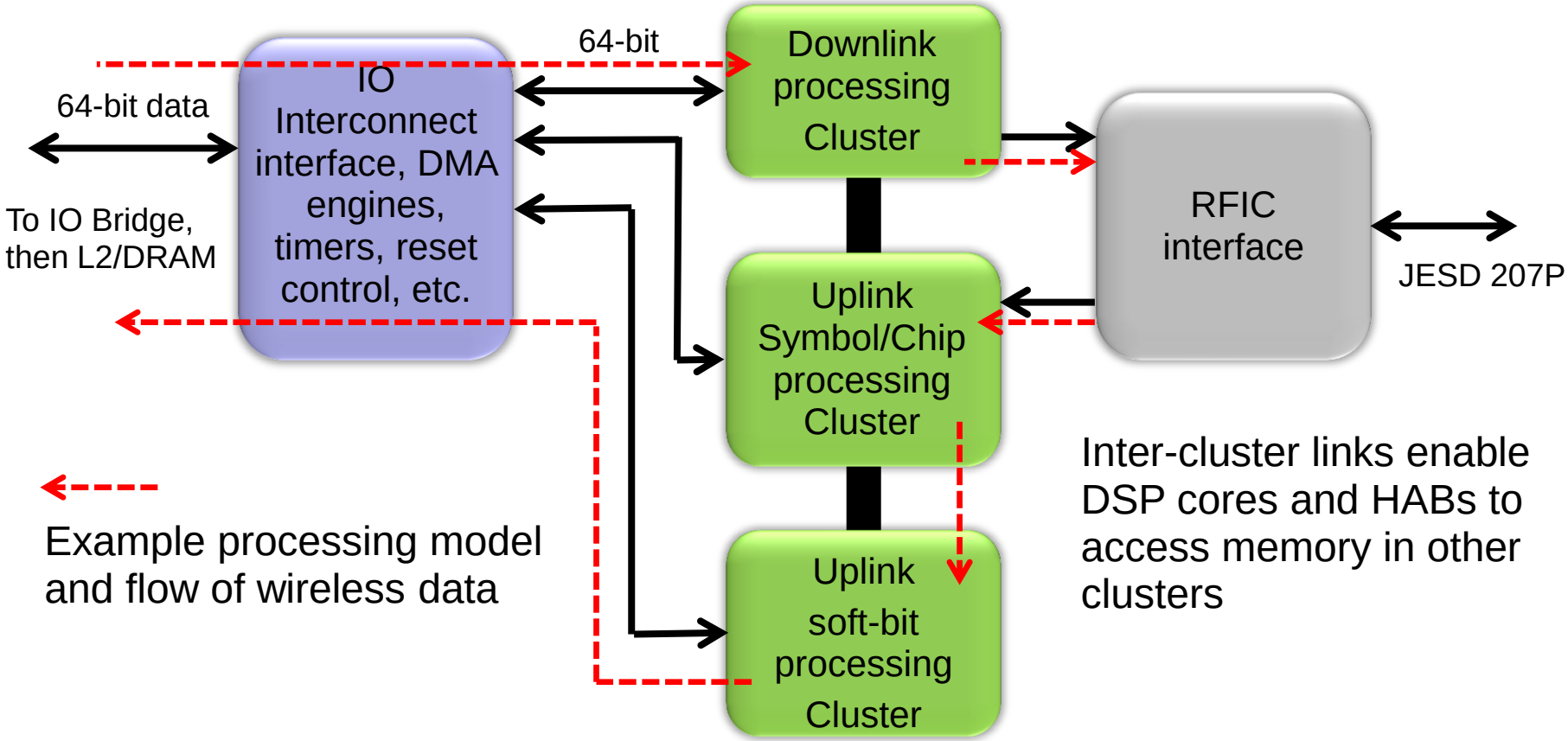
A Cluster of the Baseband Module

CAVIUM



128-bit dual load/store paths enable VLIW DSP cores to fetch two 128-bit vector operands + processing in single cycle

CNF71XX Baseband Architecture



Shared memory interconnect enables flexibility in optimizing the processing models and flows

OCTEON Multicore

CAVIUM

Wireless L2 & L3, Transport, Control, WiFi, Customer Apps

- OCTEON Fusion = OCTEON Multicore + Baseband module
- The OCTEON Multicore part of the SoC is the same architecture as OCTEON Multicore SoCs which have been widely deployed for designing base stations

CPU cores

- 4x OCTEON MIPS64 cores
- Shortest L1 and last-level-cache (L2) latencies among multicore processors
- Power optimizer™ per-core software controlled power reduction
- Fine-grained clock gating

Hardware accelerators

- Comprehensive packet processing hardware: Headers parsing, classification, RED, QoS, buffer allocation, L4 checksums, traffic rate limiting & scheduling
- Crypto, packet order, work scheduling, timers for TCP and RLC, RoHC

Low latency interconnect

- Split-transaction interconnects and L2 cache run at core frequency

OCTEON enhanced MIPS64 core

CAVIUM

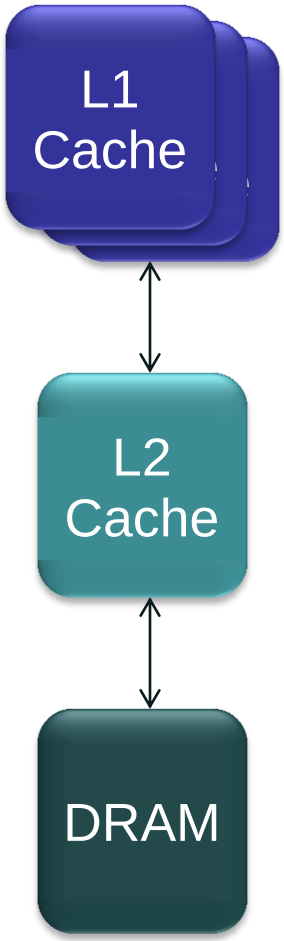
Custom designed efficient 64-bit CPU core

- Dual-issue, 8+ stages. Optimized for perf/watt, perf/area
- Short 3 cycles L1 cache load-to-use latency
- MIPS64 r3 instruction set + >80 optimizing instructions

Examples of optimizing instructions added

- Atomic memory ops (increment, add, fetch-and-add, etc.)
- Insert/extract arbitrary bit fields within a word
- Branch if certain bit field contains a set bit or not
- Compare operands and set bit0 for equal / not equal
- Additional flavors of prefetch and cache hints
- Population count
- Unaligned load/store

OCTEON Cache Policies



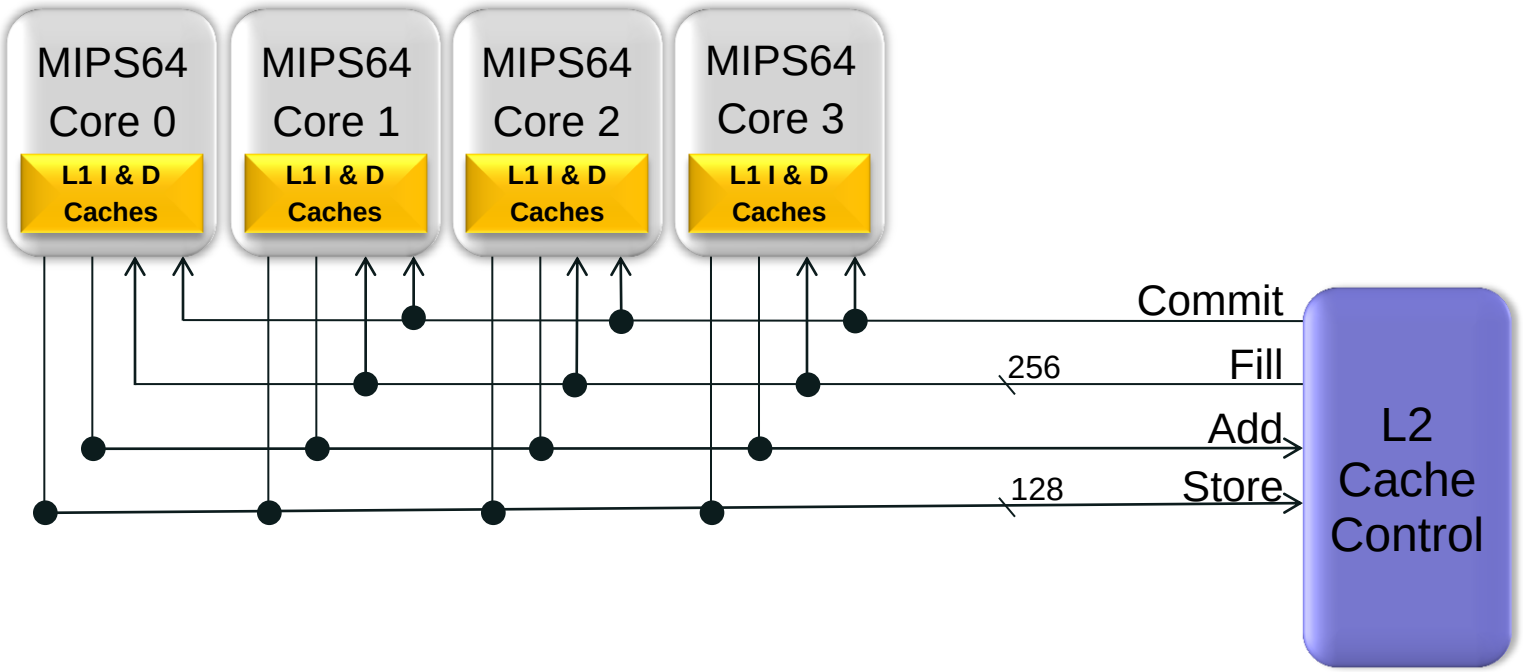
L1 <-> L2 Cache: Write-through

- Excellent performance for networking and wireless applications
- Minimal per-CPU-core cost (power, area)
- Lowest possible read latencies
- Allows many outstanding stores, optimizations
- Automatic L1 error correction

L2 Cache <-> DRAM: Write-back

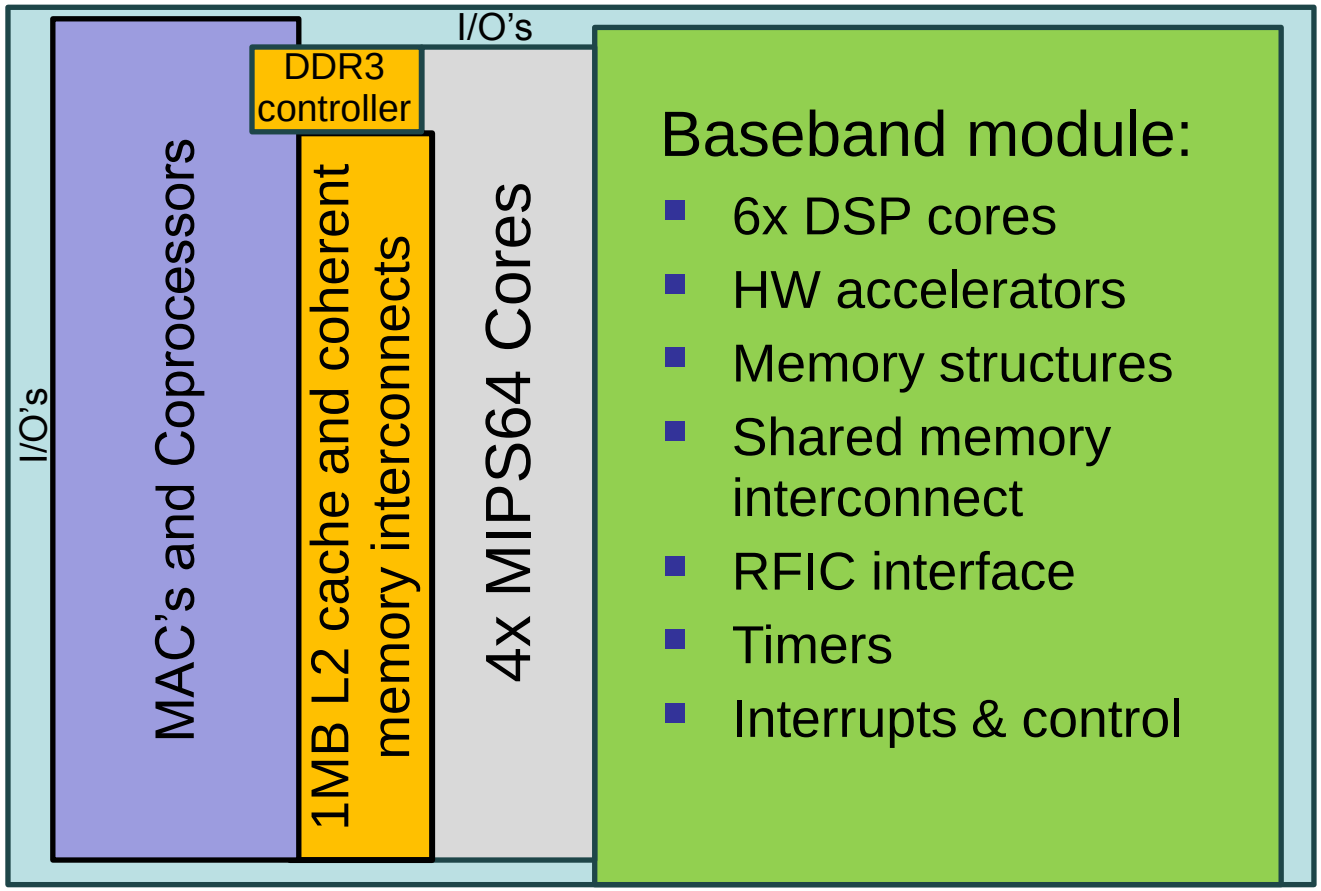
- Standard DDR3 DRAM DIMM's are highest performance with block transfers
- Minimizes required DRAM bandwidth
- Don't-write-back feature (e.g. for most of packet data) plus additional cache hints

CNF71XX Coherent Interconnect

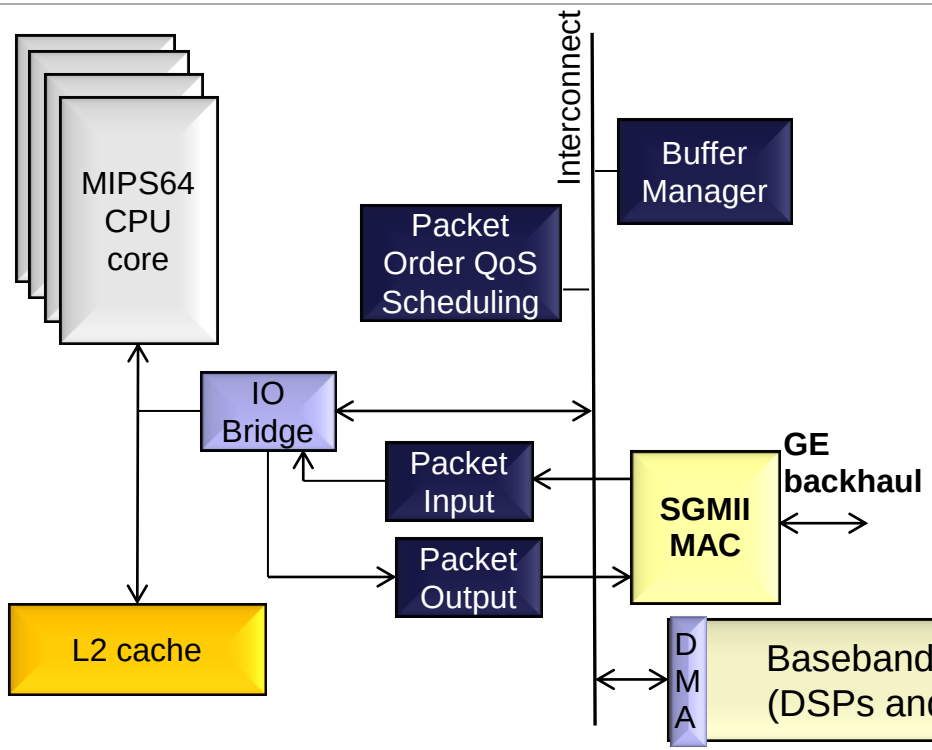


64-bit CPU cores, split-transaction interconnect, L2 cache & controller all run at core frequency

CNF71XX Chip Floorplan



Packet/Data Flow: LTE Downlink (DL) Processing



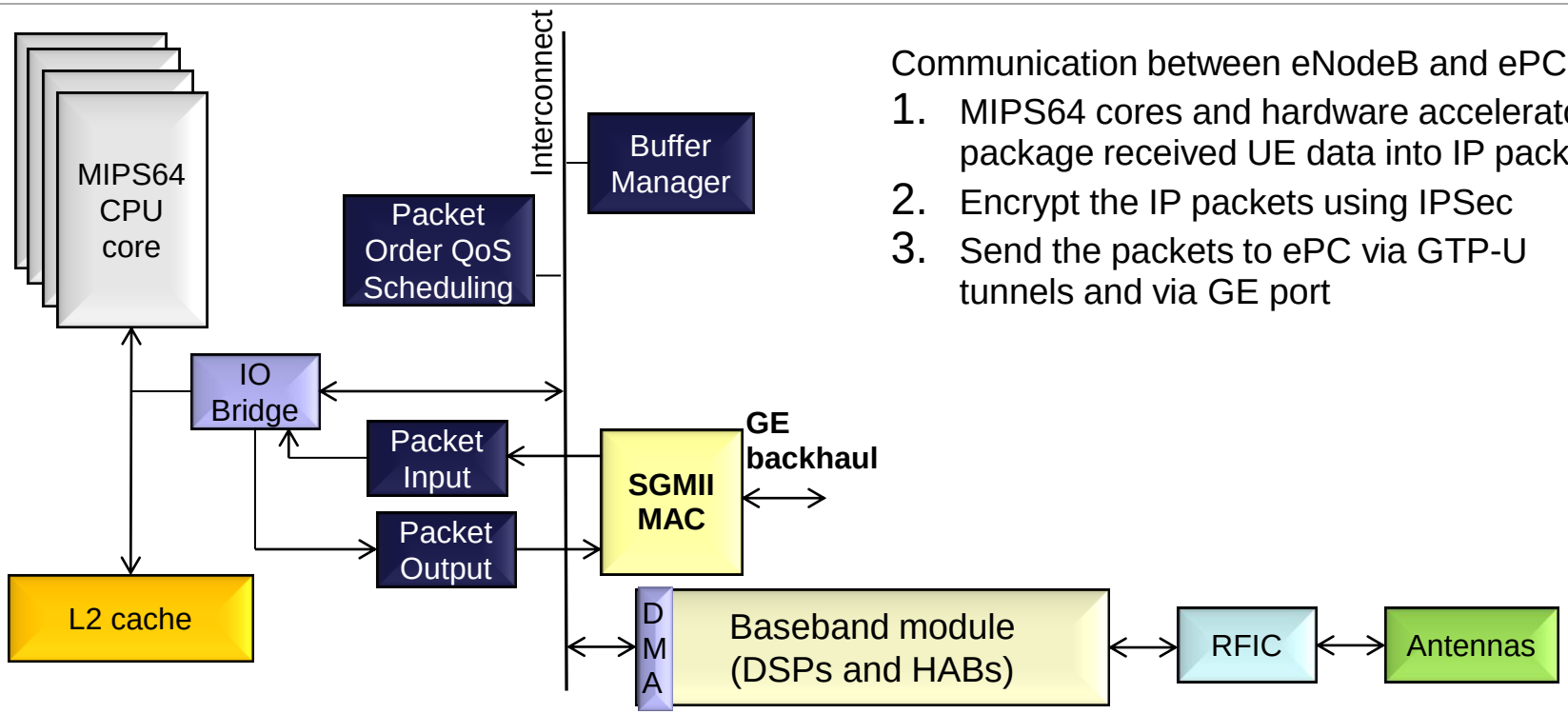
Communication between eNodeB and ePC:

1. ePC sends user packets to eNodeB over GTP-U tunnels. Packets arrive via GE port
2. Packet Input hardware handles all Ethernet packet receive, parsing headers to identify flow for packet order and QoS, allocating buffers, and DMAing packet data to buffers in L2 cache/memory
3. MIPS64 cores and hardware accelerators terminate the packet data, including IPSec decrypt

Communication between eNodeB and UE's with 1ms TTI (transmission time interval):

1. MIPS64 cores and accelerators process PDCP, RLC and MAC protocol layers.
2. MAC layer processing schedules data and wireless PHY configuration for DL transmission
3. Baseband hardware DMAs data from L2 cache to its local memory
4. Downlink DSP cores and HABs complete DL processing and transmit data out via RF interface

Packet/Data Flow: LTE Uplink (UL) Processing



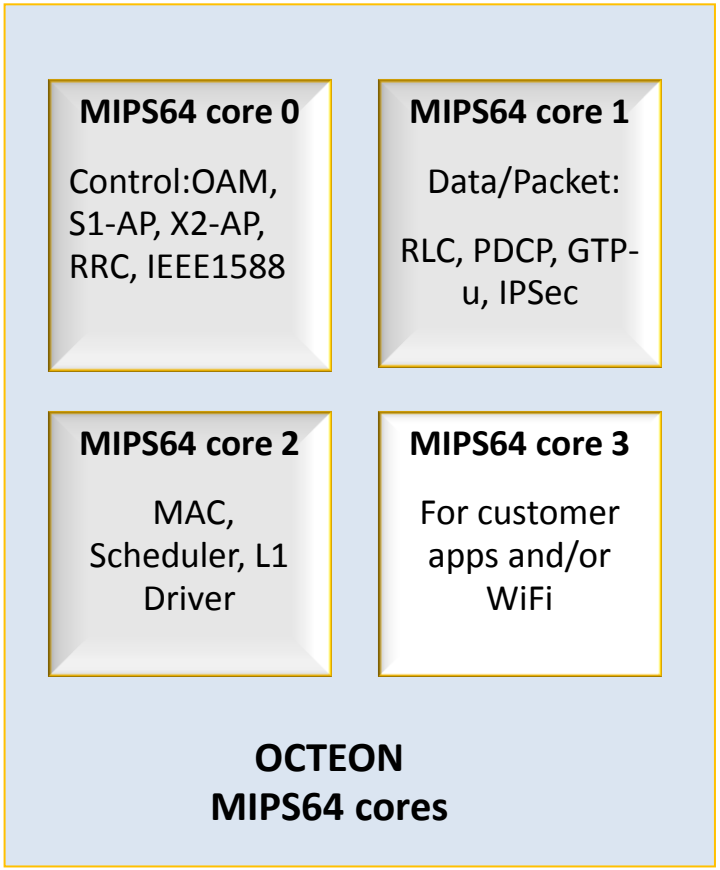
Communication between eNodeB and ePC:

1. MIPS64 cores and hardware accelerators package received UE data into IP packets
2. Encrypt the IP packets using IPSec
3. Send the packets to ePC via GTP-U tunnels and via GE port

Communication between eNodeB and UE's with 1ms TTI (transmission time interval):

1. PHY baseband processes UL traffic and detects random access from UE's
2. PHY baseband DMAs processed UL data to L2 cache
3. MIPS64 cores and accelerators process MAC, RLC, and PDCP layers to terminate received UE traffic into packets.

Mapping eNodeB to Multicore

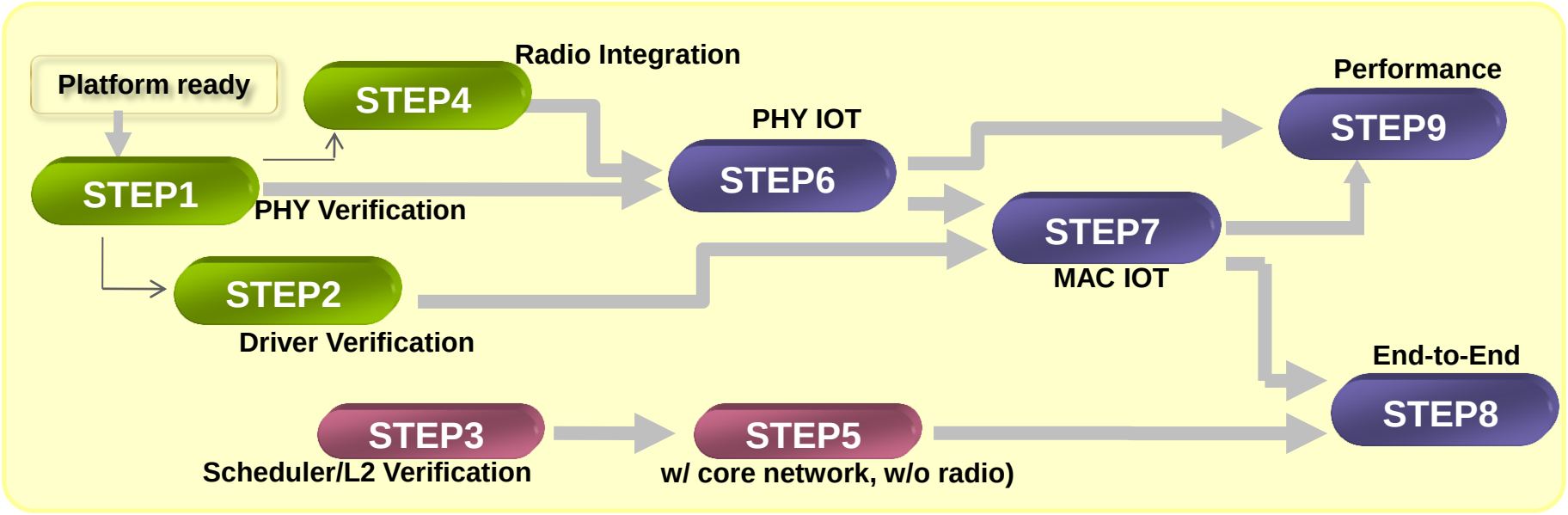


- Example partitioning : LTE eNodeB AP
 - MAC and L1 driver on one core
 - Easy to meet LTE 1ms TTI
 - Quick response to PHY interrupts
 - RLC, PDCP, Transport on one core
 - Option to partition L2 cache to avoid cache pollution from control processing
 - Control processing on one core
 - 1 core free
 - Headroom for WiFi and service provider applications
- Small Cell Forum API compliant

Quad-core delivers required headroom and deterministic performance for real-time LTE and other processing

CNF71xx Complete End-to-end Validation

- STEP1 – PHY + Driver S/W + PLT (Physical Layer Test)
- STEP2 – PHY + Driver S/W + Scheduler
- STEP3 – L1 + L2 + L3
- STEP4 – PHY + Modem + Radio
- STEP5 – Core network + Basestation (L2/L3 stacks, S1 I/F)
- STEP6 – IOT (Interoperability Testing) in PHY (PLT + Modem + Radio + UE L1)
- STEP7 – IOT in MAC (w/ UE L1/L2)
- STEP8 – IOT in E2E (w/ UE over full protocol stacks)
- STEP9 – DL/UL Performance Measurements w/ UE



Summary

- OCTEON Fusion CNF71XX
 - High performance “base station on a chip” SoC
 - LTE 20MHz, 150Mbps DL + 150Mbps UL, 2x2 MIMO, 128 users
 - OCTEON Fusion = OCTEON multicore + baseband
 - Same OCTEON software for small to macro cells
 - End-to-end interoperability and performance verified
- Optimized for Base station designs
 - Delivers deterministic real-time performance, low power, and high integration, with significant compute headroom
 - 4x enhanced & efficient 64-bit (OCTEON MIPS) CPU cores
 - 6x Baseband optimized DSP vector processors
 - Many hardware accelerators
 - Optimized for short latencies and deterministic performance

Backup

Cavium: Company Summary



- Founded 2001
- NASDAQ IPO (CAVM) 2007
- Locations: US, India, China, TW
- 2011 Revenues : $259M, +26% YOY
- 5 year CAGR: ~50%
- Profitable with Strong Financials, Zero Debt

- Addressing Multi-billion dollar Networking, Communications, Storage and Digital Home markets.
- MIPS64 and ARM based Multi-core Processor SoCs; Multi-core Search and Security Processors
- All Top Networking, Wireless and Security Vendors use Cavium

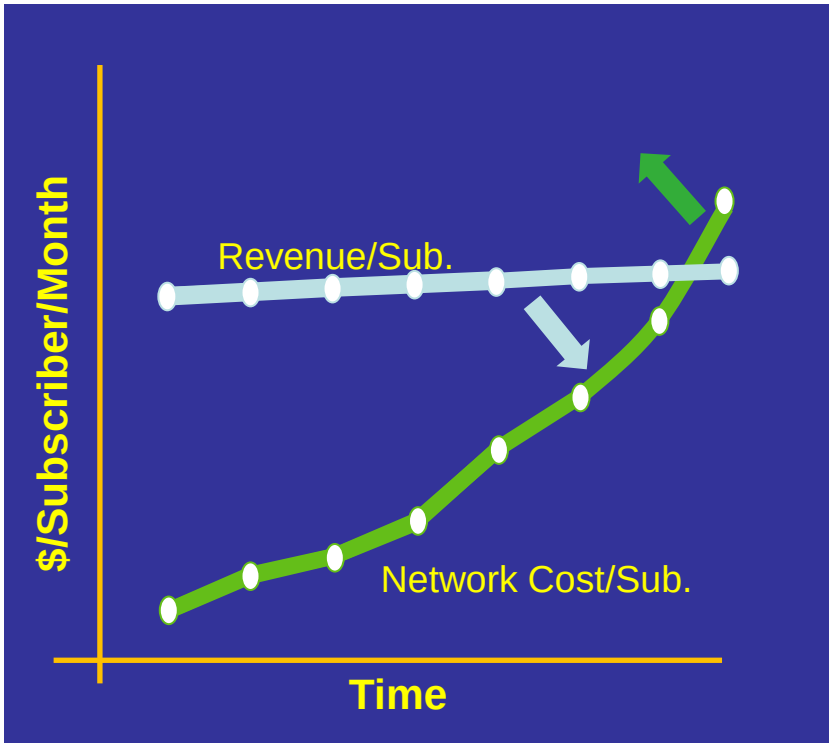
Carriers coping with 1000x traffic increase and no extra revenue

Smart devices multiply traffic



* Monthly basic mobile phone data traffic

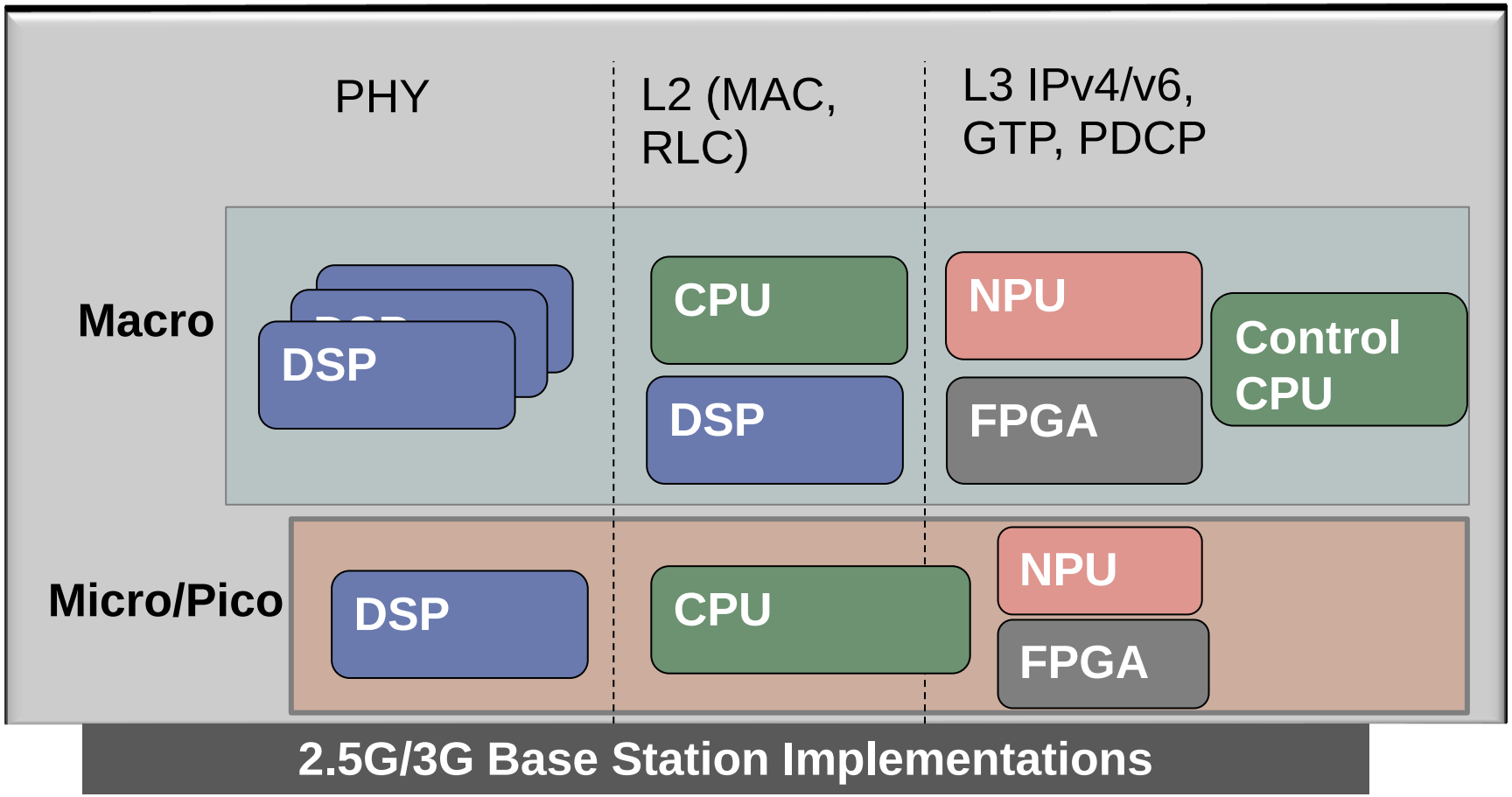
Source: Cisco VNI Mobile, 2011



Heterogeneous Radio Access Network

- Macro base stations are expensive (CAPEX and OPEX)
- Augment Macro with Small cell base stations to add capacity and coverage cost effectively

Previous Generation Base Stations



Before Multi-core SoCs became available, Base Station designs required many components, microcode programming on NPU, general purpose CPUs, FPGAs, and many development environments. High complexity



FSM™ Femtocell Station Modem

A Highly Integrated, Performance Driven Chipset for the Small Cell Market

Luca Blessent

Notices

Copyright © 2012 QUALCOMM Incorporated.
All rights reserved.

QUALCOMM is a registered trademark of QUALCOMM Incorporated in the United States and may be registered in other countries. Other product and brand names may be trademarks or registered trademarks of their respective owners.

Qualcomm reserves the right to make changes to the product(s) or information contained herein without notice. No liability is assumed for any damages arising directly or indirectly by their use or application. The information provided in this document is provided on an “as is” basis.

Outline

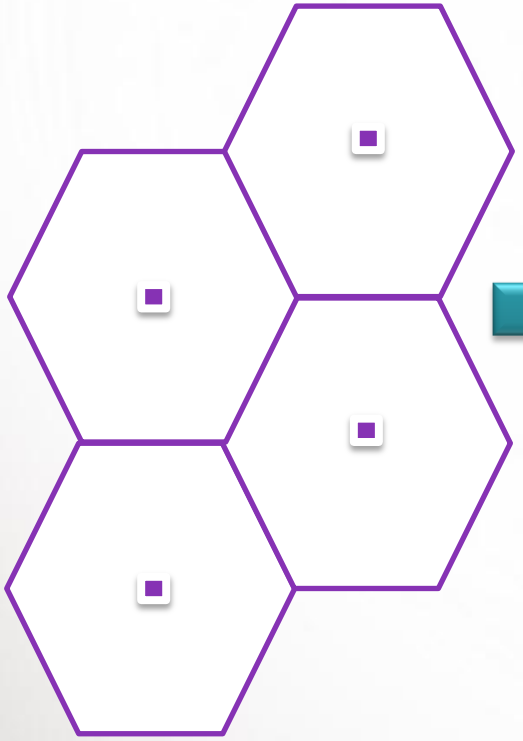
- Small Cells: Motivation and Implications
- Cellular Access Point Evolution
- The FSM9xxx Chipset
- Design Challenges
- Selected Advanced Features
- FSM9xxx Based Access Point
- Power Consumption
- Summary and Closing Remarks

Copyright © 2012 QUALCOMM, Incorporated.

Traditional Cellular Coverage Model

+ Small Cells

New Cellular Topology



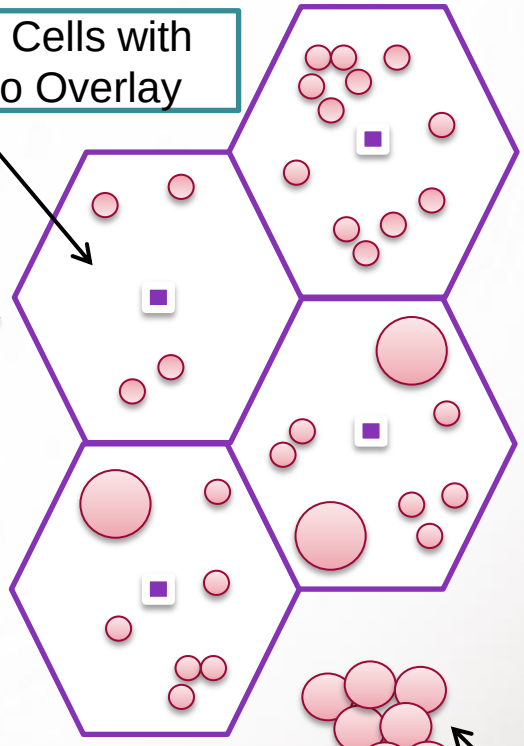
Data Demand ↑

Capacity ↑

Limited Spectrum

Improved User Experience

Small Cells with Macro Overlay



Interference Management

Low Cost

Low Power

Advanced Features





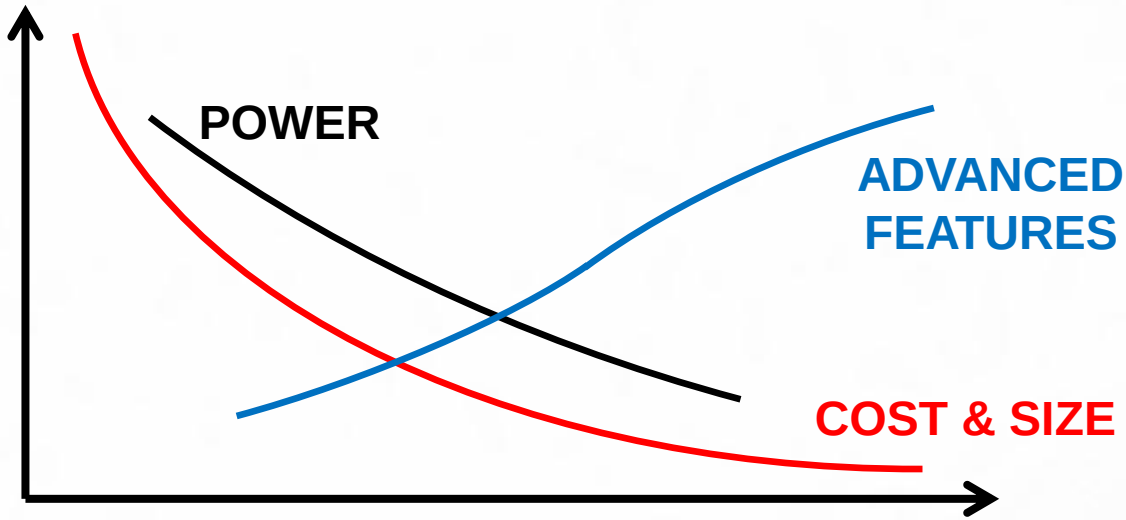
Ultra Compact Small Cell Access Point

Neighborhood Femtocells

New Requirements for Cellular Access Points

Copyright © 2012 QUALCOMM, Incorporated.

Cellular Access Point Evolution

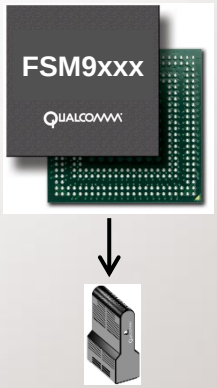












Copyright © 2012 QUALCOMM, Incorporated.

The FSM9xxx SoC

Chip Layout



Key Stats

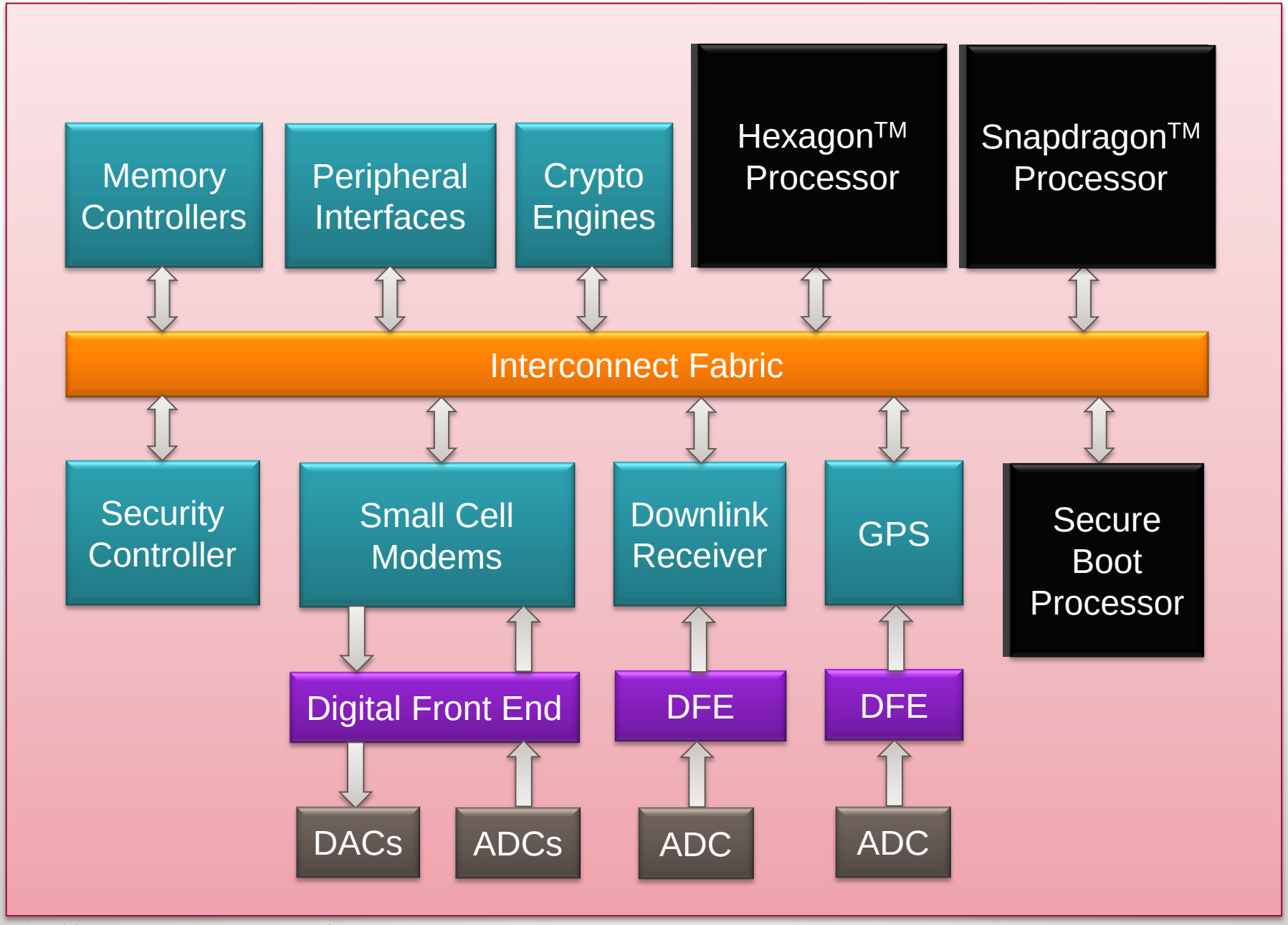
- 45 nm
- ~ 1.8 W for realistic full load
- Sampling commercially since April 2011

Key Features

- Small Cell Modem
- Integrated GPS
- Snapdragon™ Application Processor
- Security provisions
- Interference management

Copyright © 2012 QUALCOMM, Incorporated.

The FSM9xxx Architecture



Copyright © 2012 QUALCOMM, Incorporated.

Processors

Snapdragon™ Processor

- Qualcomm's 1st generation CPU, codenamed "Scorpion"
- 1 GHz
- ARMv7 ISA
- ~ 1.6x DMIPS/MHz w.r.t. ARM11
- Optimized for low power
- Open processor
- Handles L3, OA&M, etc.

Hexagon™ Processor

- Qualcomm's custom DSP
- 600 MHz
- Multi-threaded
- Closed processor
- Handles L1 hardware control and L2

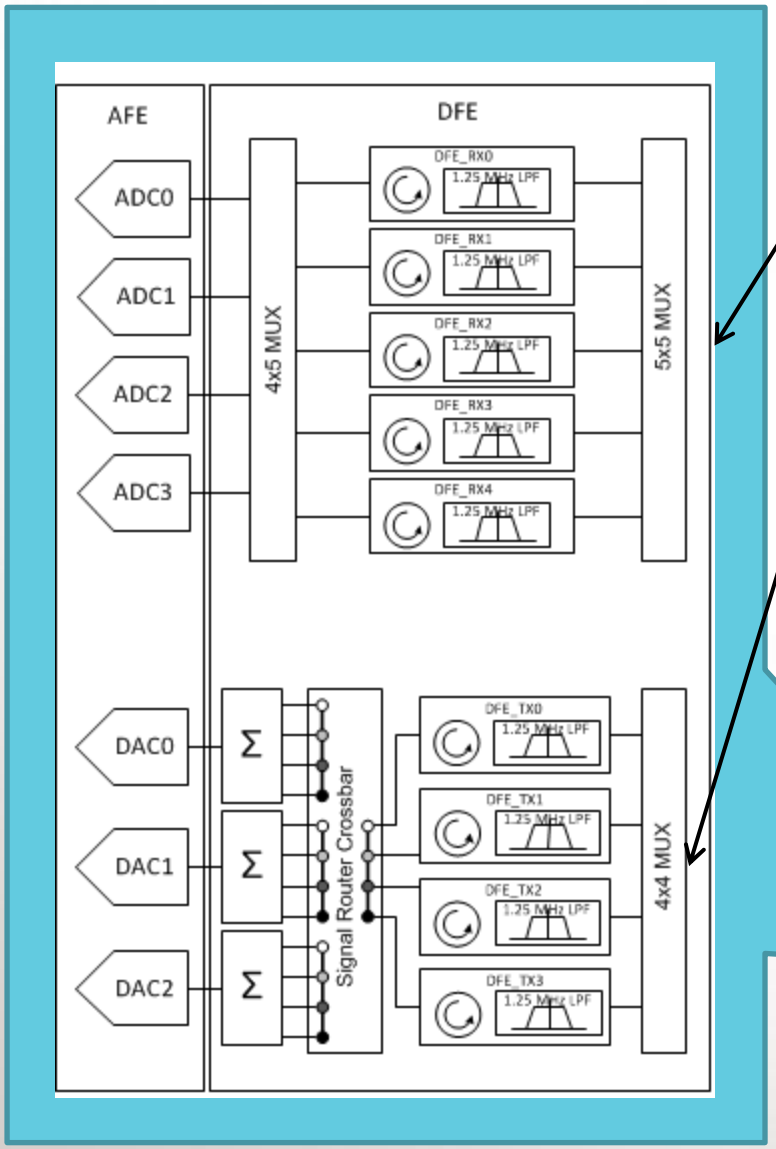
Copyright © 2012 QUALCOMM, Incorporated.

Design Challenges

- Need to combine base station and mobile functionality
 - Downlink processing for neighbor discovery and self-configuration
- Aggressive power consumption target
 - < 5W for full solution
- Stringent security requirements for residential deployment
 - Requires on-chip trusted execution environment
- Uncompromised modem performance
 - Up to 16 Multi-RAB UMTS users
 - 28 Mb/s downlink throughput
 - 5.7 Mb/s uplink throughput
 - Rx and Tx diversity
- Support for advanced interference management features
 - Additional processing chains for beaconing and uplink measurements

Copyright © 2012 QUALCOMM, Incorporated.

Advanced Signal Processing

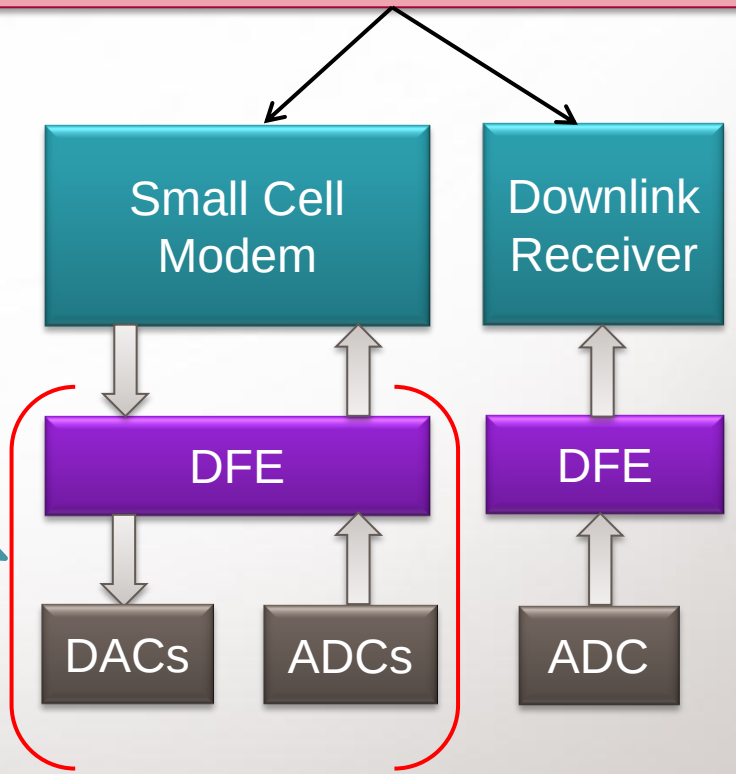


Additional processing chains:

1) Downlink beaconing to facilitate system reselection
2) Uplink mobile and interference sensing

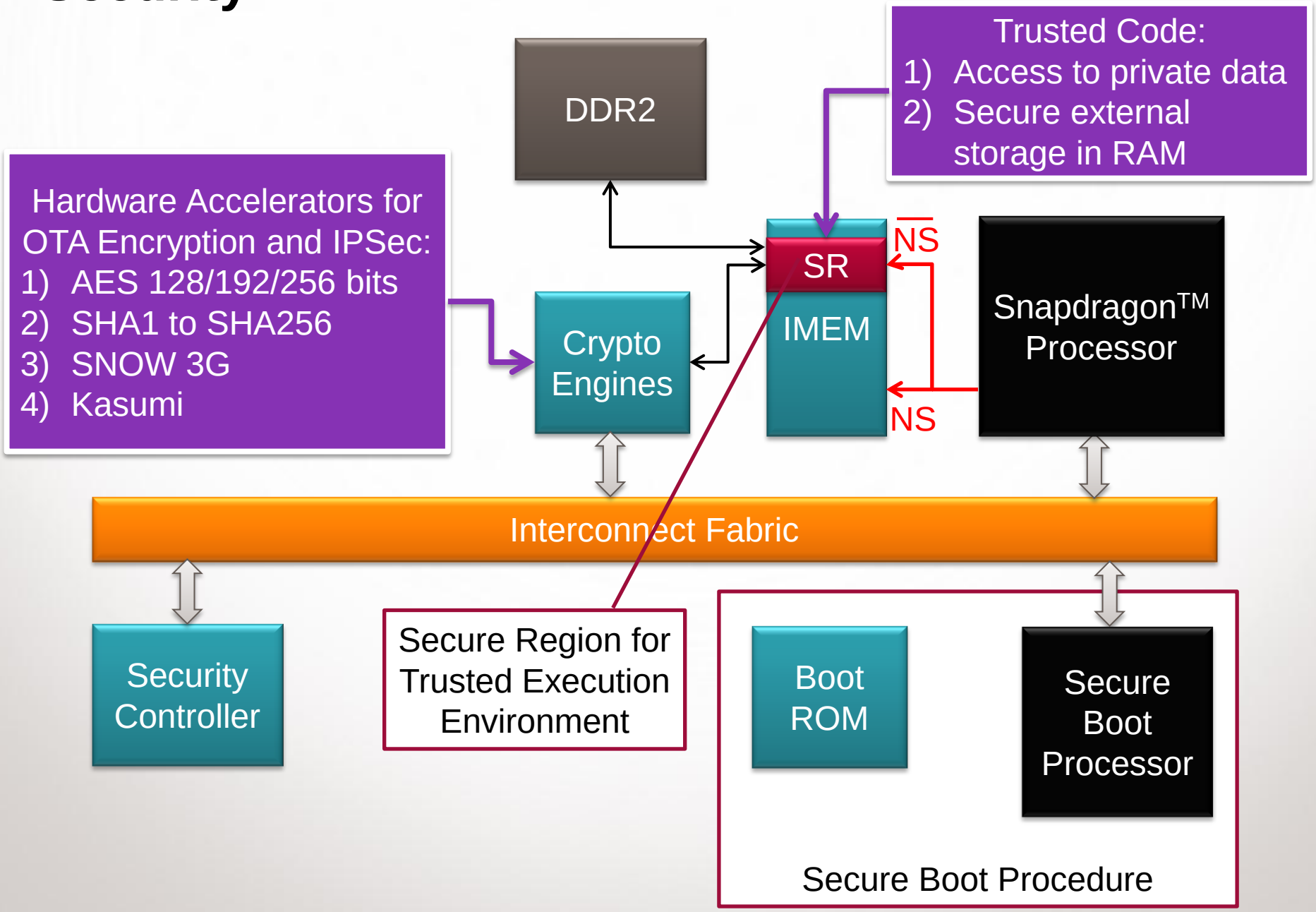
Simultaneous small cell service and downlink sniffing:

1) Dynamic interference management
2) Continuous VCTCXO disciplining



Copyright © 2012 QUALCOMM, Incorporated.

Security



Copyright © 2012 QUALCOMM, Incorporated.

The FSM9xxx Chipset

FSM9xxx Baseband Processor

- FSM92xx SKUs for UMTS
- FSM98xx SKUs for CDMA2000



FTR8700 Transceiver

- 2x2 wideband (25 MHz) chains
- Global UMTS and CDMA2000 bands

RTR8605 Receiver

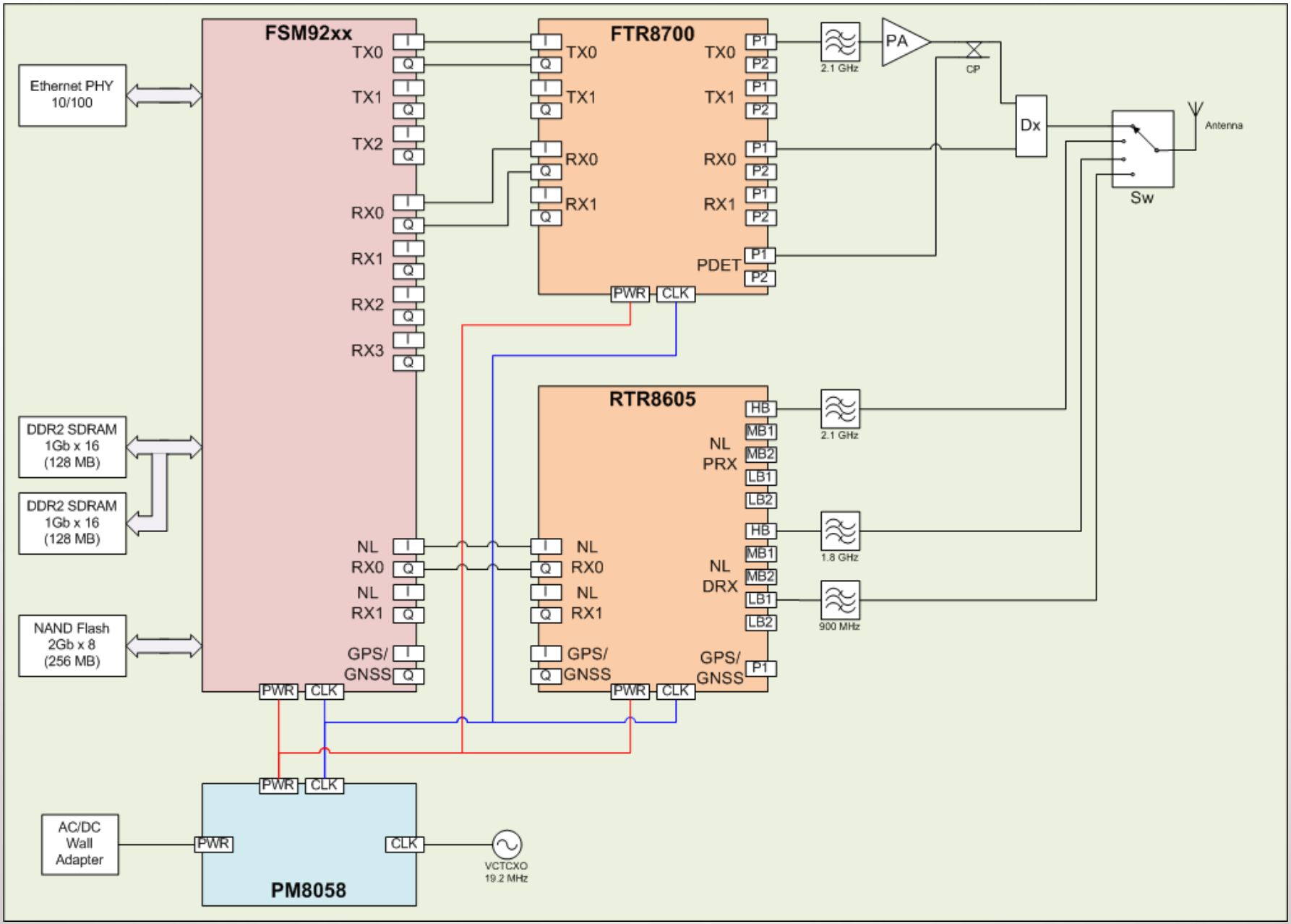
- Downlink receiver
- GPS receiver

Power Management IC

- Voltage regulators
- System clocks

Copyright © 2012 QUALCOMM, Incorporated.

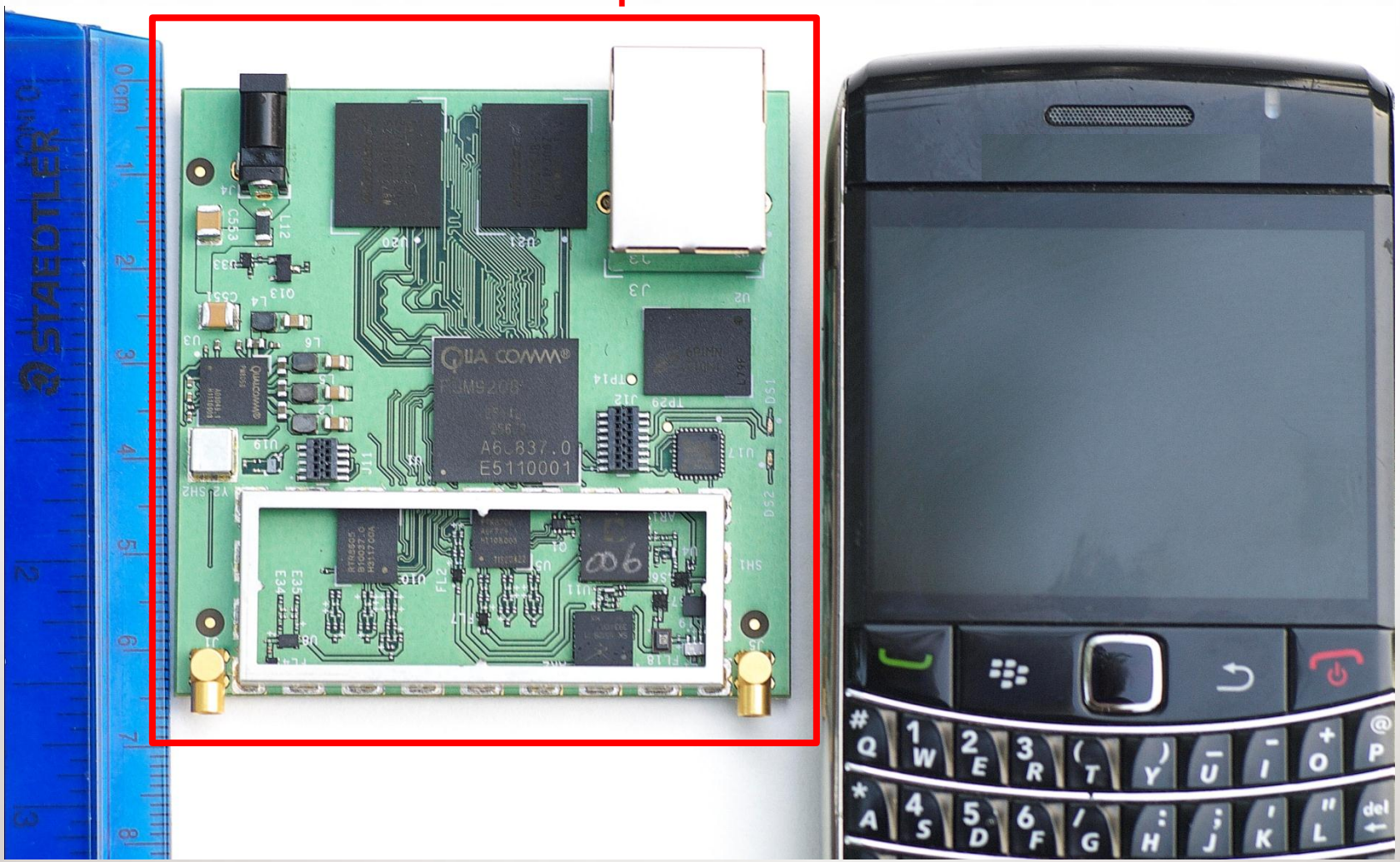
FSM9xxx Based AP: Functional View



Copyright © 2012 QUALCOMM, Incorporated.

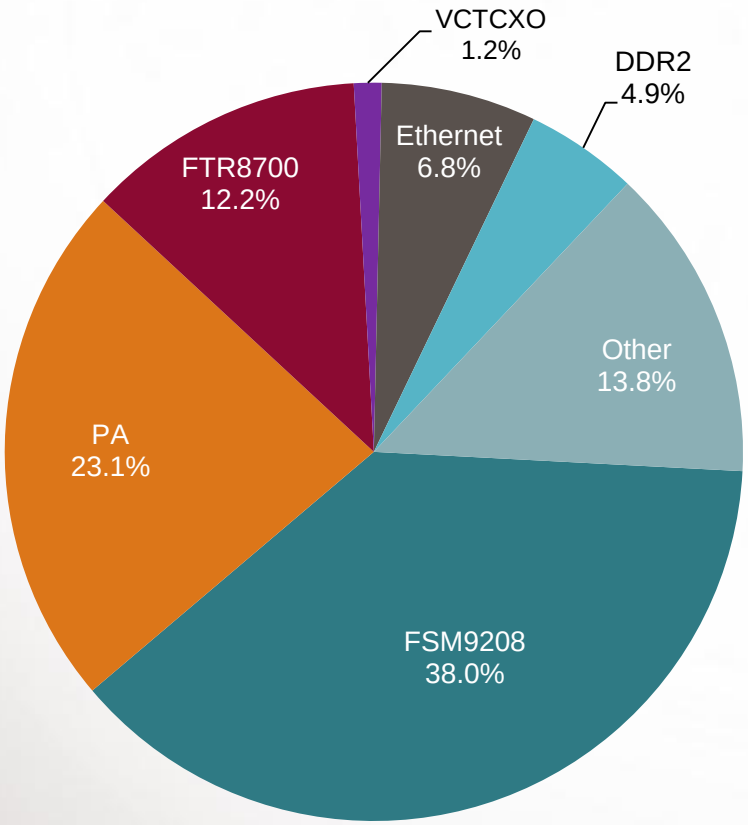
FSM9xxx Based AP: Implementation

2.5 in. x 2.5 in., 6-layer Board
Small Cell Access Point Implementation



Copyright © 2012 QUALCOMM, Incorporated.

Power Consumption



Total AP Power: 4.8 W

Test Configuration:

- 8 user residential femtocell (FSM9208)
- HSDPA + EUL operation
- 1.9 GHz band
- 13 dBm maximum Tx power
- Single Tx/Rx
- GPS and downlink receiver active
- Measurements at room temperature

Copyright © 2012 QUALCOMM, Incorporated.

Summary and Closing Remarks

- Data demand, capacity limits and economics are driving operators towards small cells
- Small cells deployment models create new opportunities and introduce new design challenges
- The FSM SoC provides a set of advanced features for improved system performance
- This SoC enables a very compact, low power small cell AP design
- The FSM9xxx chipset is Qualcomm's 1st generation small cell solution, focused on 3G
- This chipset is part of a portfolio of solutions that will include LTE, integrated Wi-Fi, and small cells evolution

Copyright © 2012 QUALCOMM, Incorporated.

QUALCOMM®

Thank You!



Medfield Smartphone SOC Intel® Atom™ Z2460 Processor

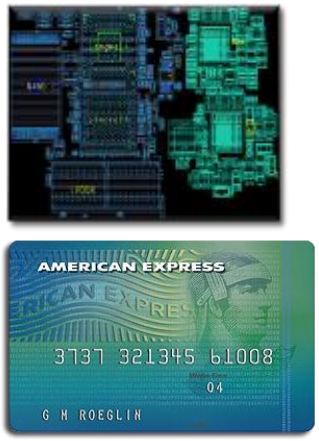
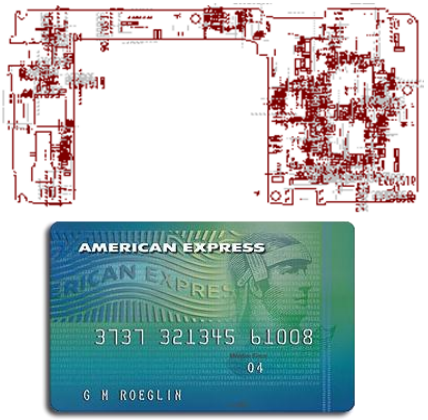


Rumi Zahir
Intel Corporation

Outline

- Low power platform progression
- Medfield platform for the Smartphone form-factor
 - Constraints, Ingredients, Package
- Penwell SOC
 - Block Diagram
 - Intel Atom™ CPU power management
 - SOC power management
 - Power management software architecture
- Medfield reference platform
- Smartphone roadmap

Low Power Platform Progression

	Moorestown (45nm)	Medfield (32nm)	
			
Board size	5,000mm2	4,150mm2	(↓ 17%)
Standby power	21mW	14mW	(↓ 33%)
Browsing power	1.2W	0.85W	(↓ 29%)
Video	+ 720p encode	+ 1080p encode	
Camera	5 mega-pixel	up to 16 mega-pixel	
Graphics	800 MPPS	2,000 MPPS	(↑ 250%)

intel

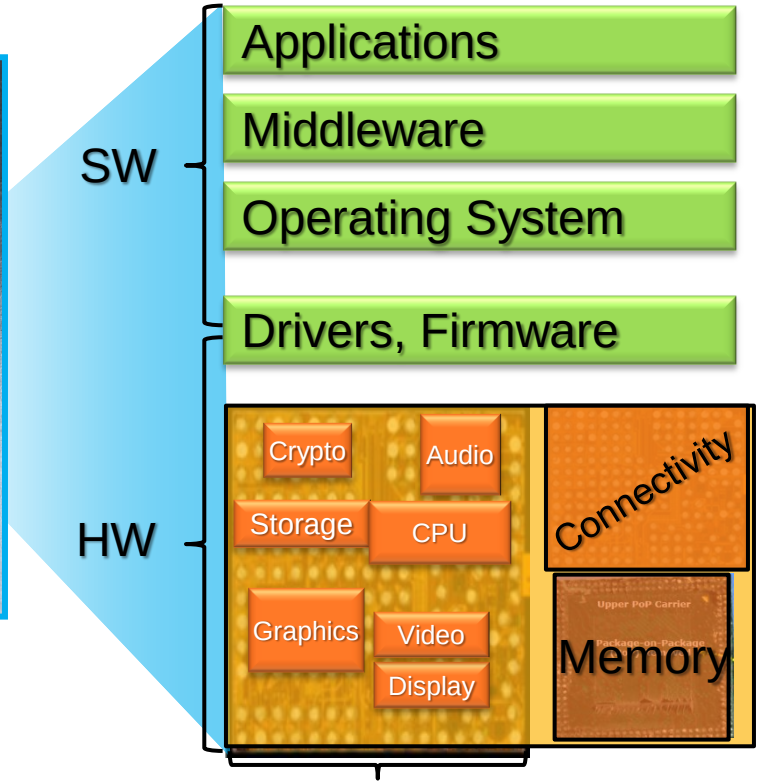
Smartphone through the Systems Lens



Intel Form Factor
Reference Design



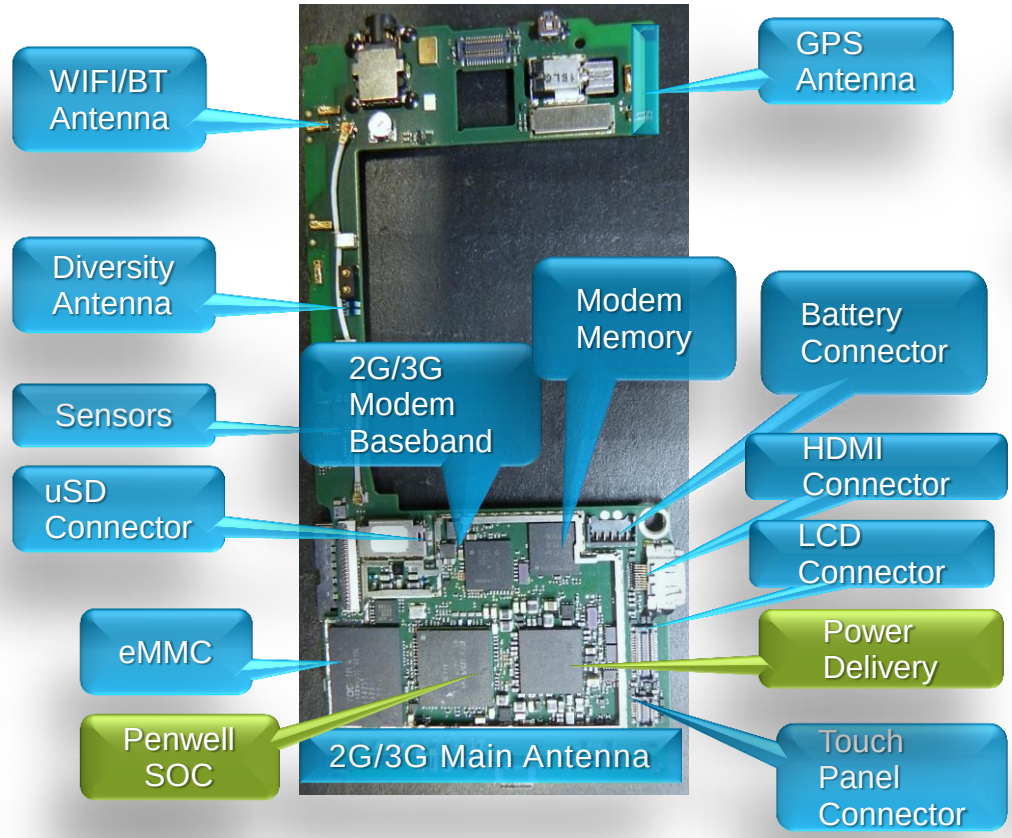
Board Design

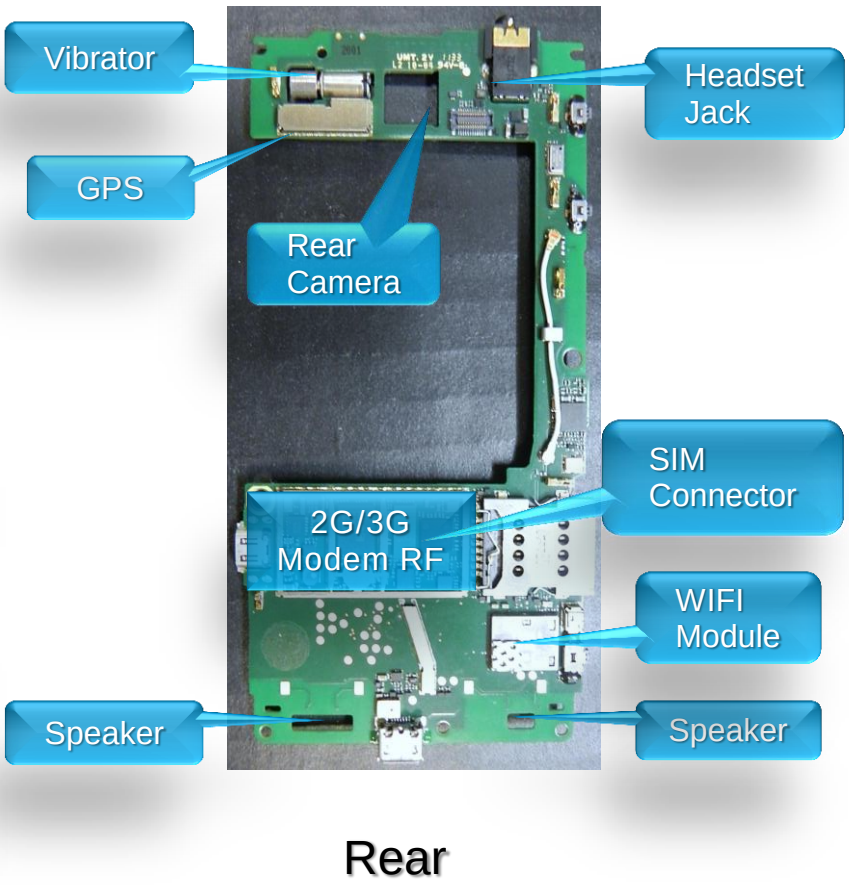


System-on-A-Chip
Integration

Design to meet Smartphone cost/power/performance requirements

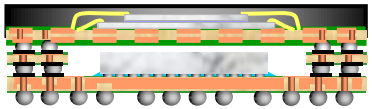
Medfield System Ingredients

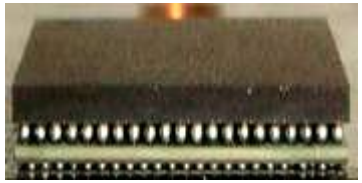




intel

Penwell SoC Package Size

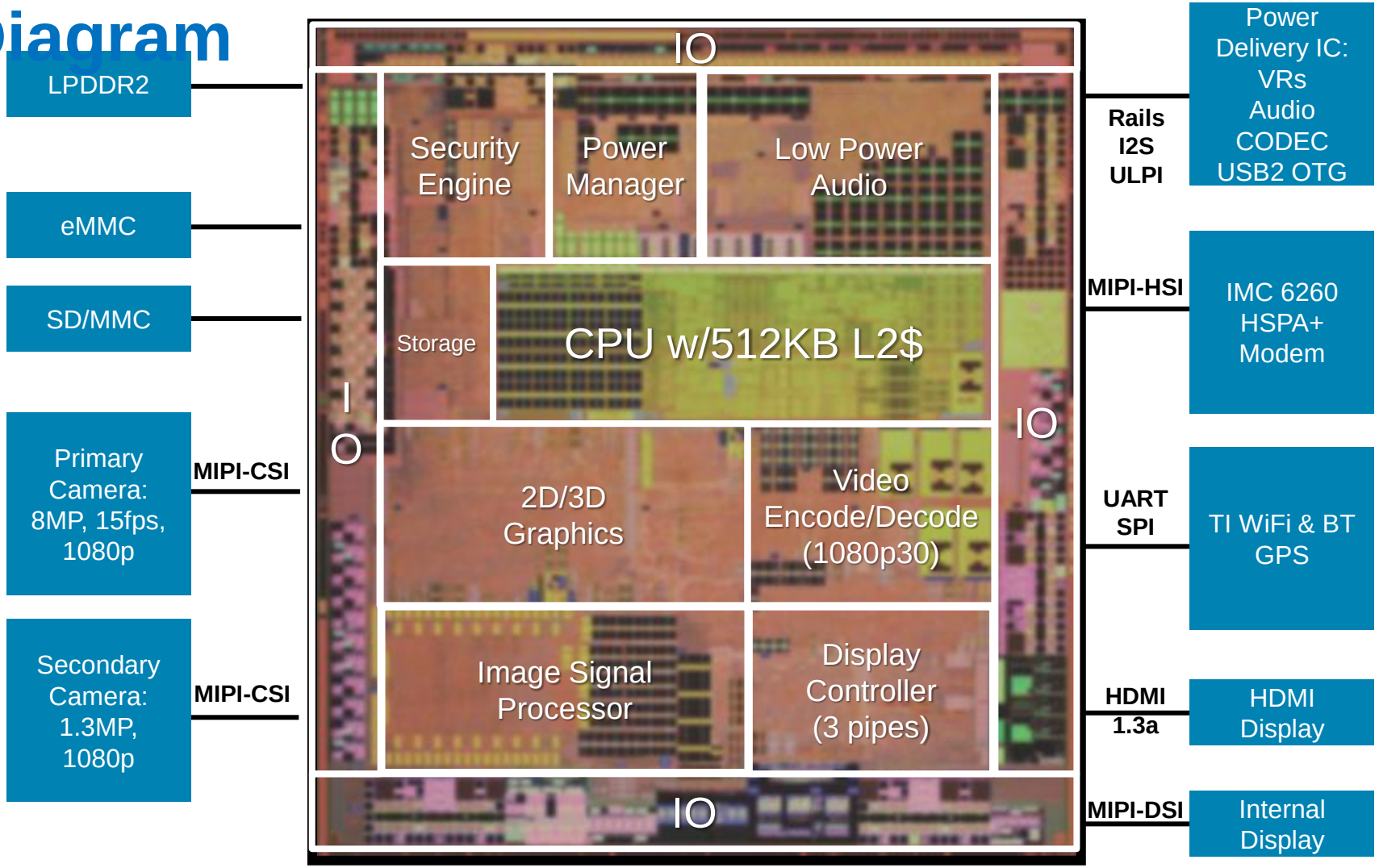




Package-on-Package (POP)
- 12 x 12 mm PoP FCMB4 – 32nm
- Non PoP SoC < 0.8 mm
- PoP z height < 1.4mm
- OEM/ODM can solder up to 2 GB of LPDDR2 memory on top of SOC

- **Memory Peak Bandwidth**
 - ✓ 6.4GB/s @ 800MT/s
 - ✓ Channels and ranks
- **Dual 32 bit channels**
 - ✓ Supports 1 or 2 ranks per channel
- **Memory Size and Density**
 - ✓ Supports total memory size of 128MB, 256MB, 512MB and 1GB per channel
 - ✓ Supports 1Gb, 2Gb and 4Gb chip densities
- **Other Features**
 - ✓ Aggressive power management to reduce power consumption
 - ✓ Proactive page closing policies to close unused pages
 - ✓ Supports different physical mappings of bank addresses to optimize for performance

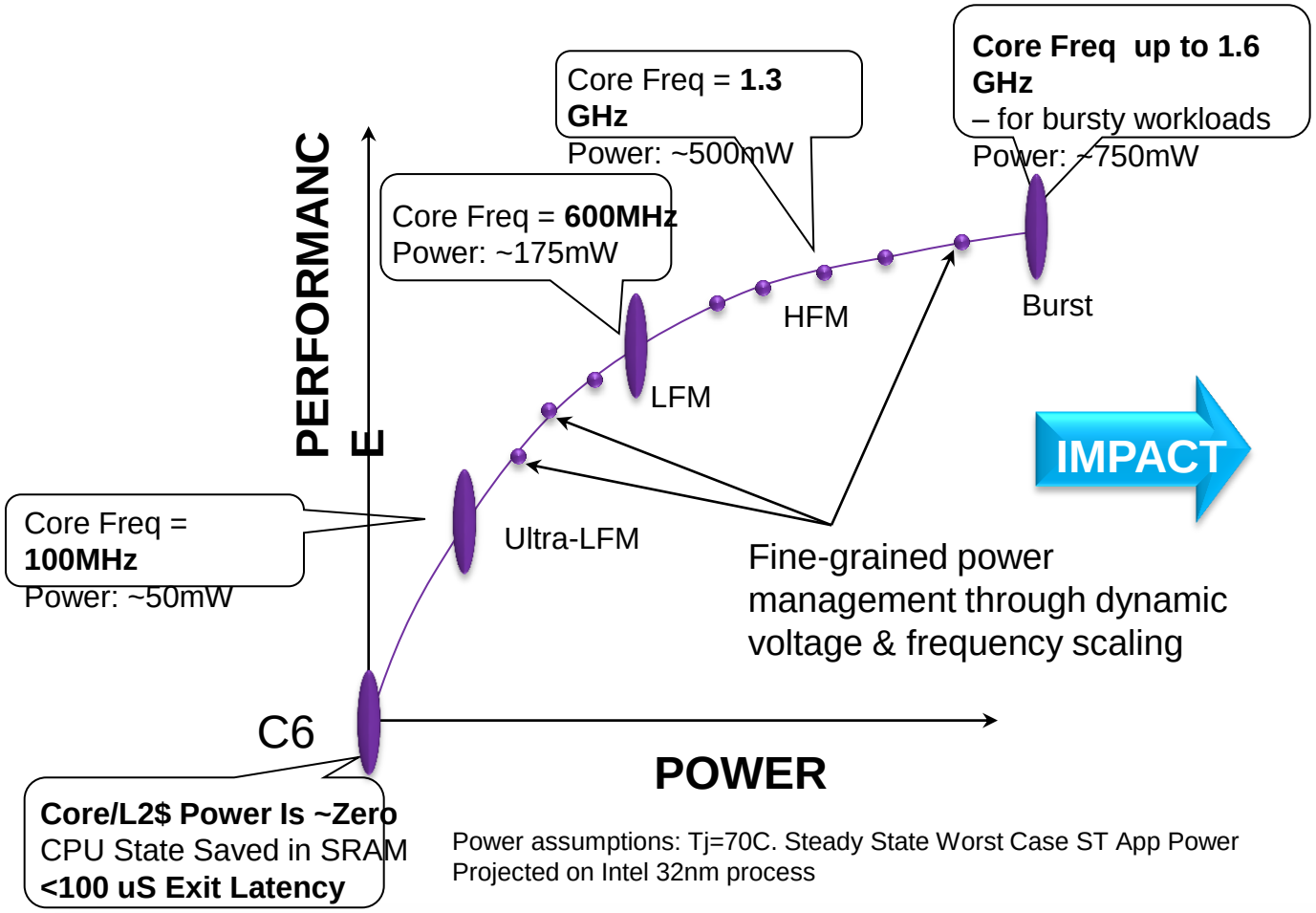
Medfield with Penwell SOC Block Diagram



Penwell SOC (Intel Hi-K 32nm Process Technology)

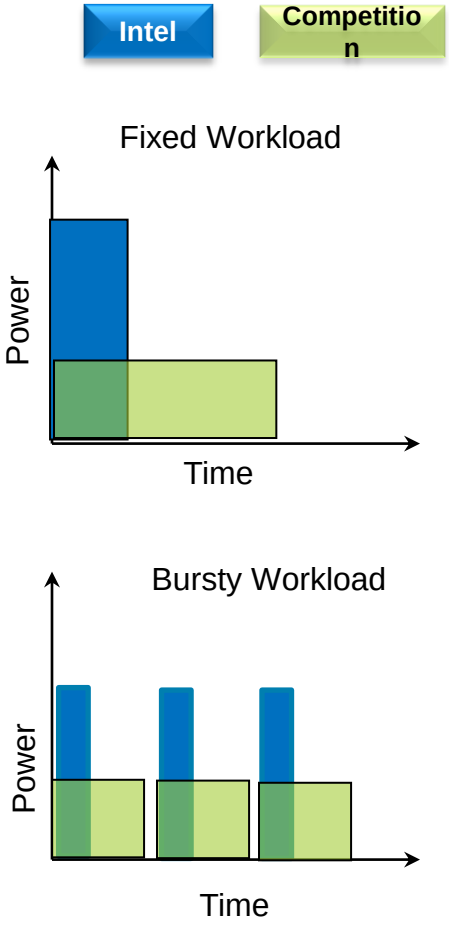
intel

Penwell CPU Dynamic Range



Power assumptions: Tj=70C. Steady State Worst Case ST App Power Projected on Intel 32nm process

IMPACT



Wide Dynamic Range & Fast Exit Latencies = Big Energy Savings

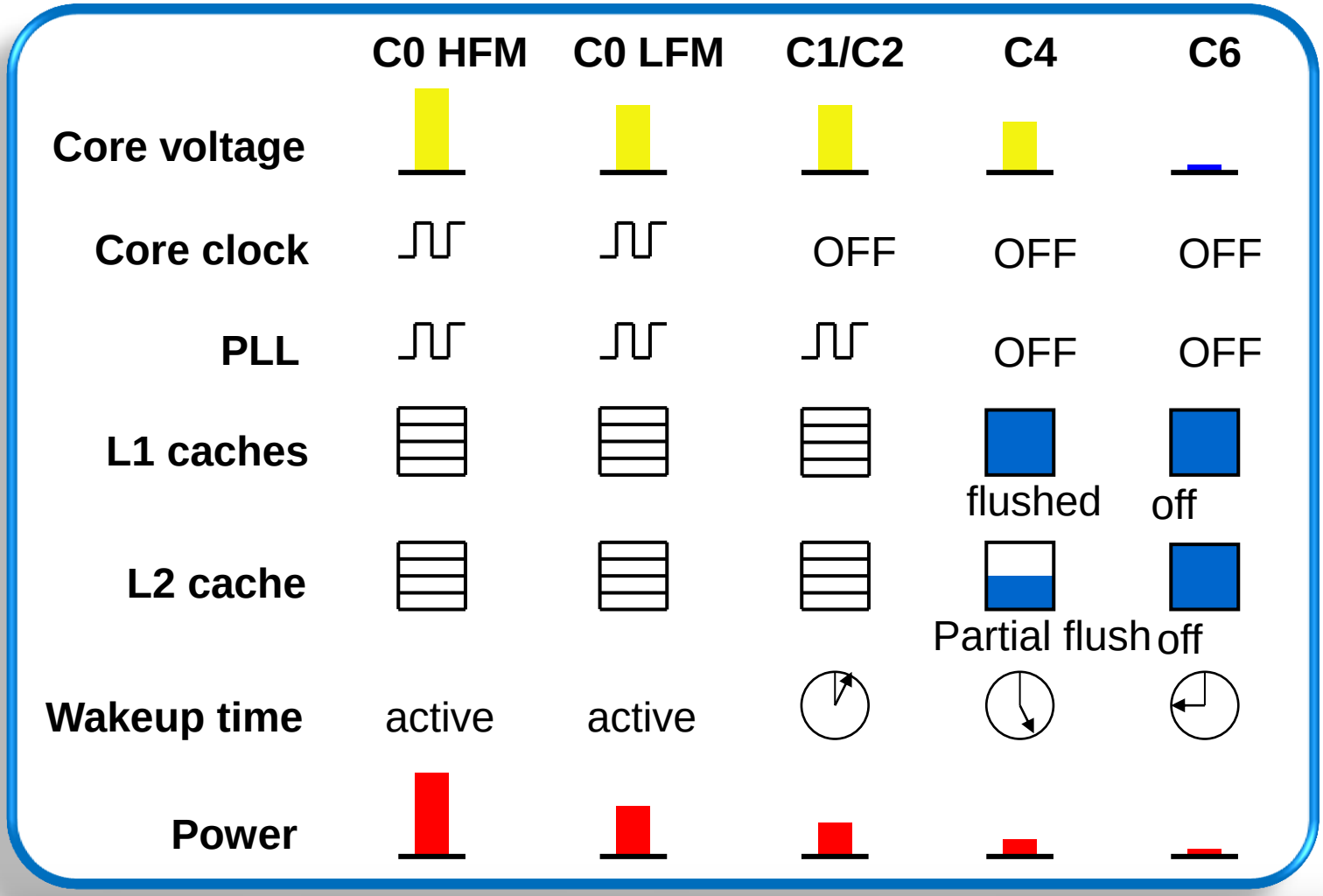
Browser Results Summary

P0	600 MHz	900 MHz	1,500 MHz
Frequency	1x	1.5x	2.5x
Performance	1x	1.41x	2.24x
Power	1x	1.29x	1.81x
Energy	1x	0.92x	0.81x

"Race to Idle" at higher frequency uses more power, but is lower energy

intel

Power C-States



The OS Is Responsible For Identifying When The Processor Needs To Be In A Certain C State And Requests The Processor To Enter That State

intel

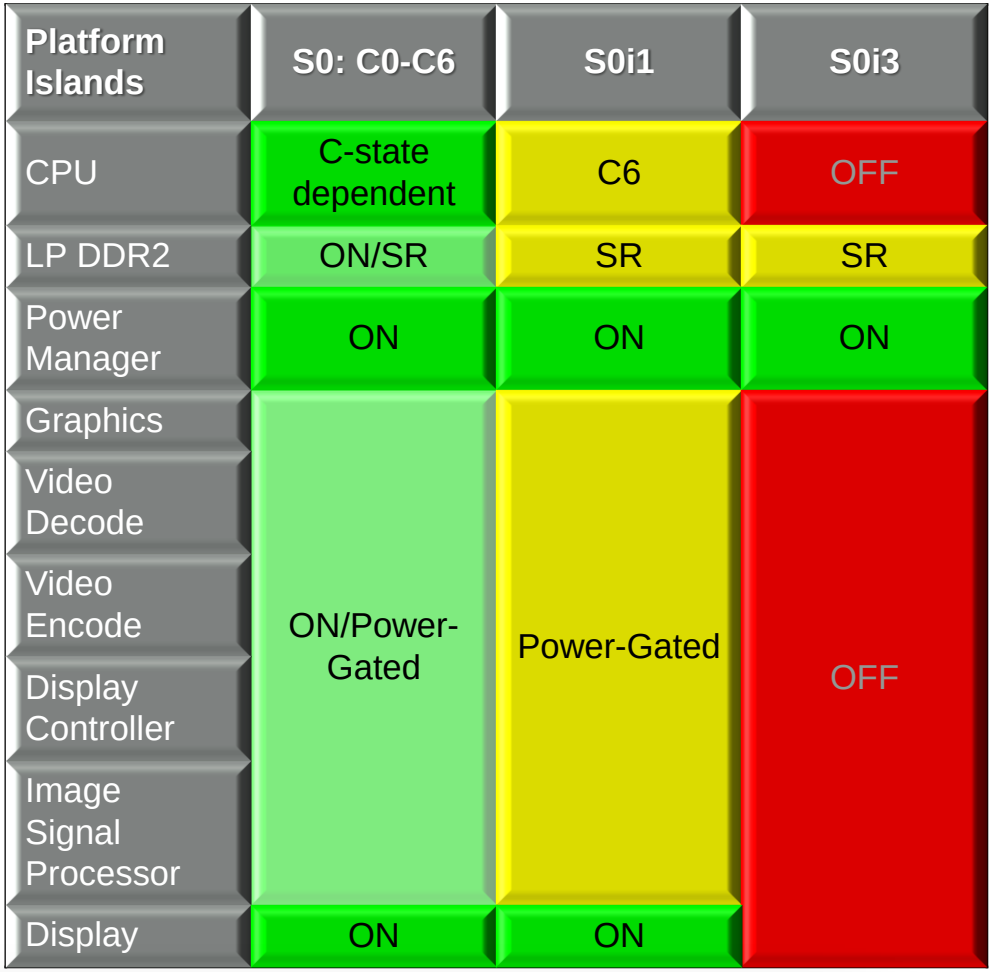
New Platform Level: “S” Ultra Low Power States

S0i1

- Used during idle (e.g. home screen, web browsing)
- Ultra Low Power: mW
- Entry-Exit Latency: µs

S0i3 / S3

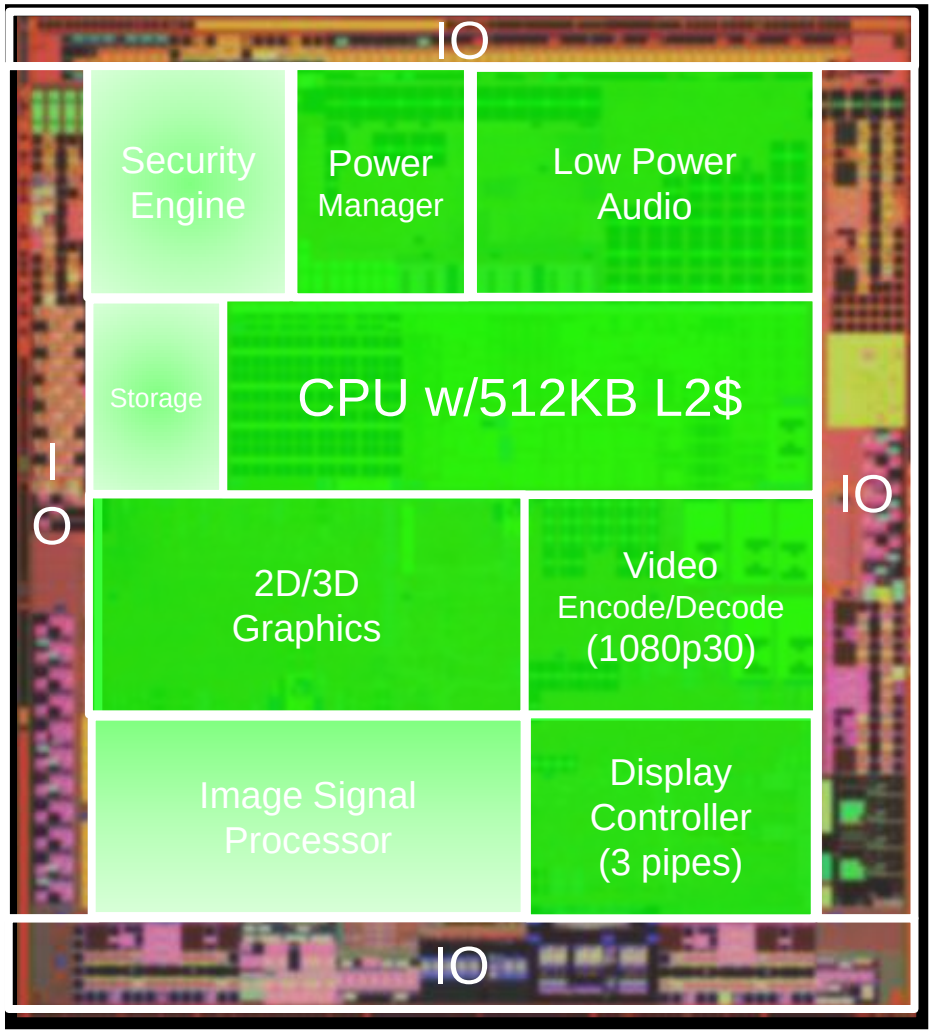
- Used when NOT interacting with the device (e.g. standby mode)
- SoC power: µW
- Entry-Exit Latency: ms

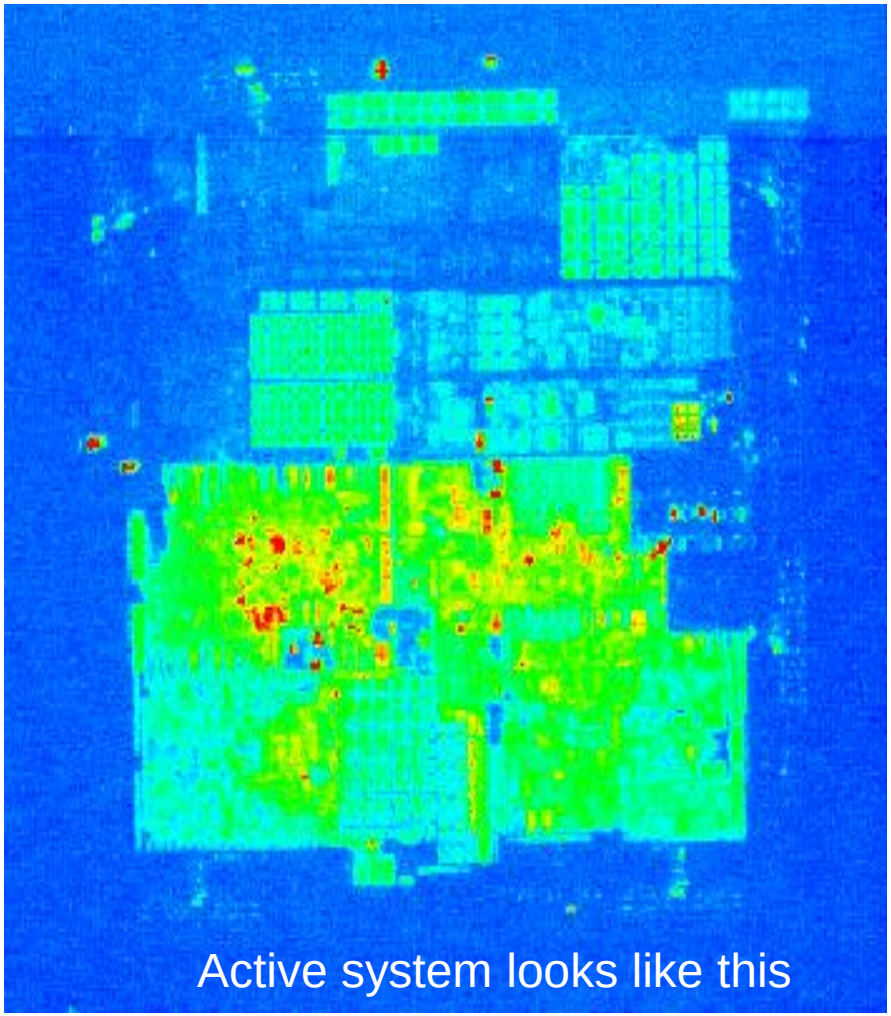


Achieves Ultra Low Power States with Best-in-Class Latency

intel

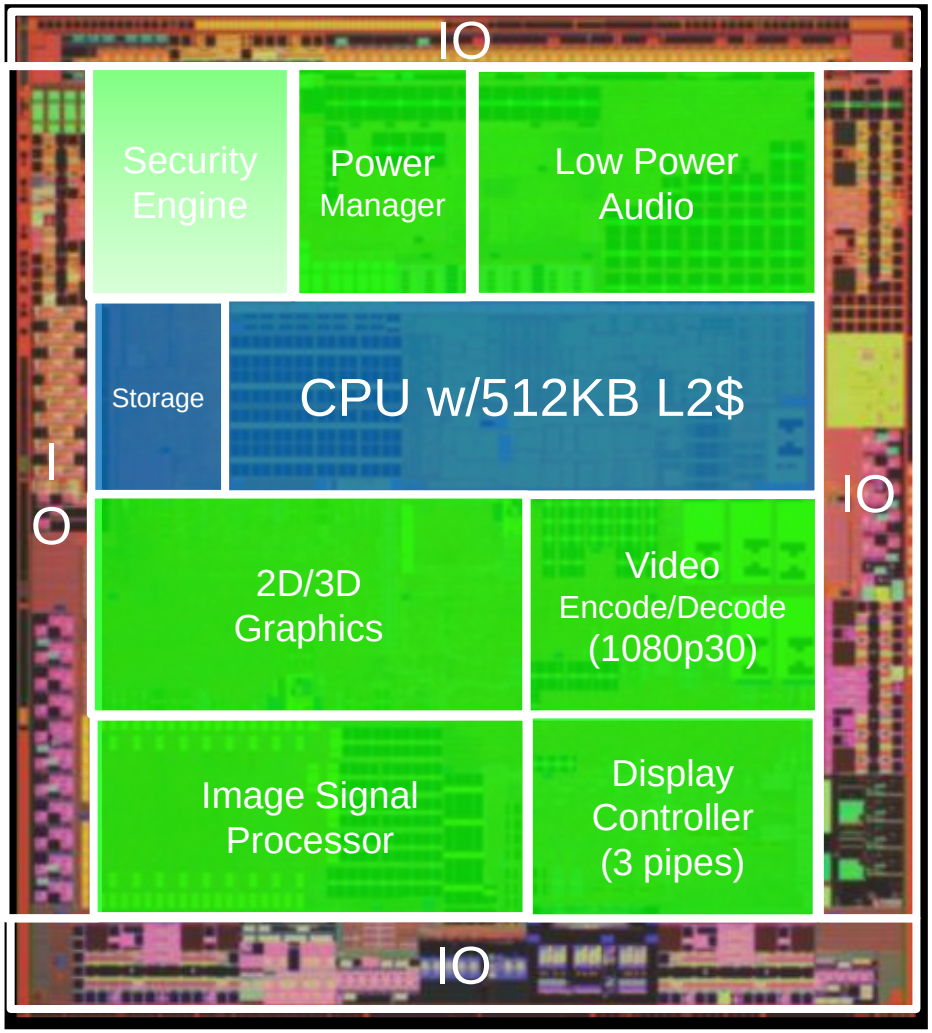
CPU Active

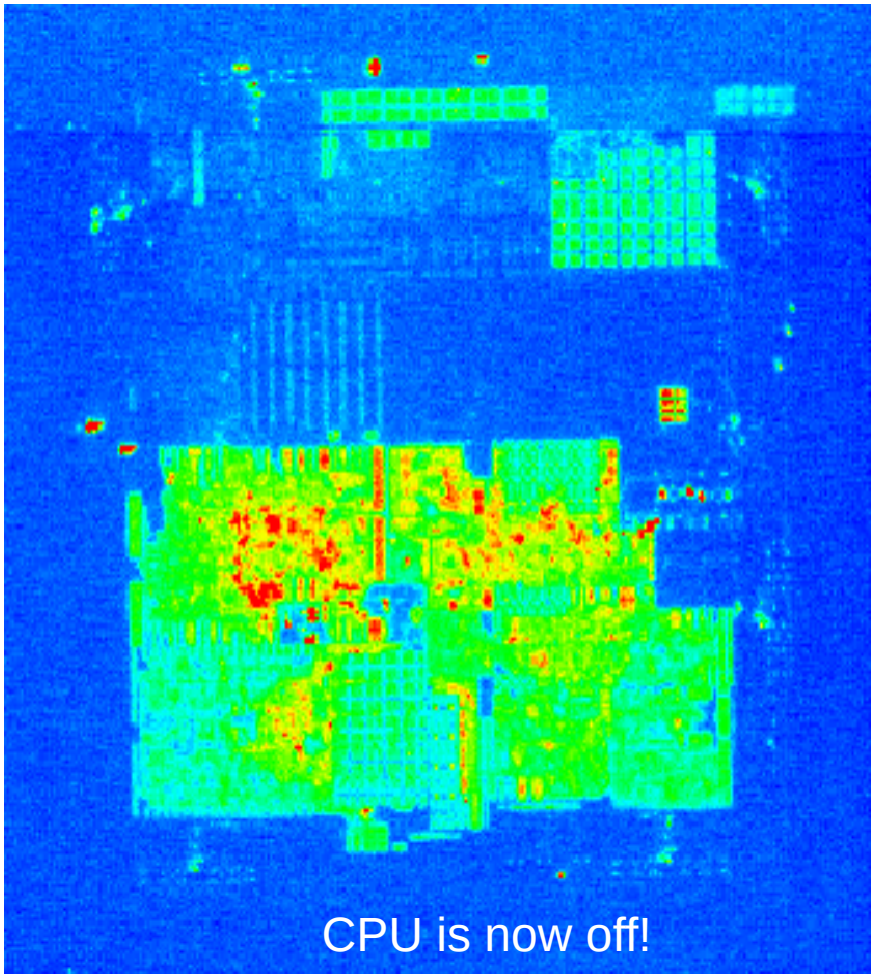




intel

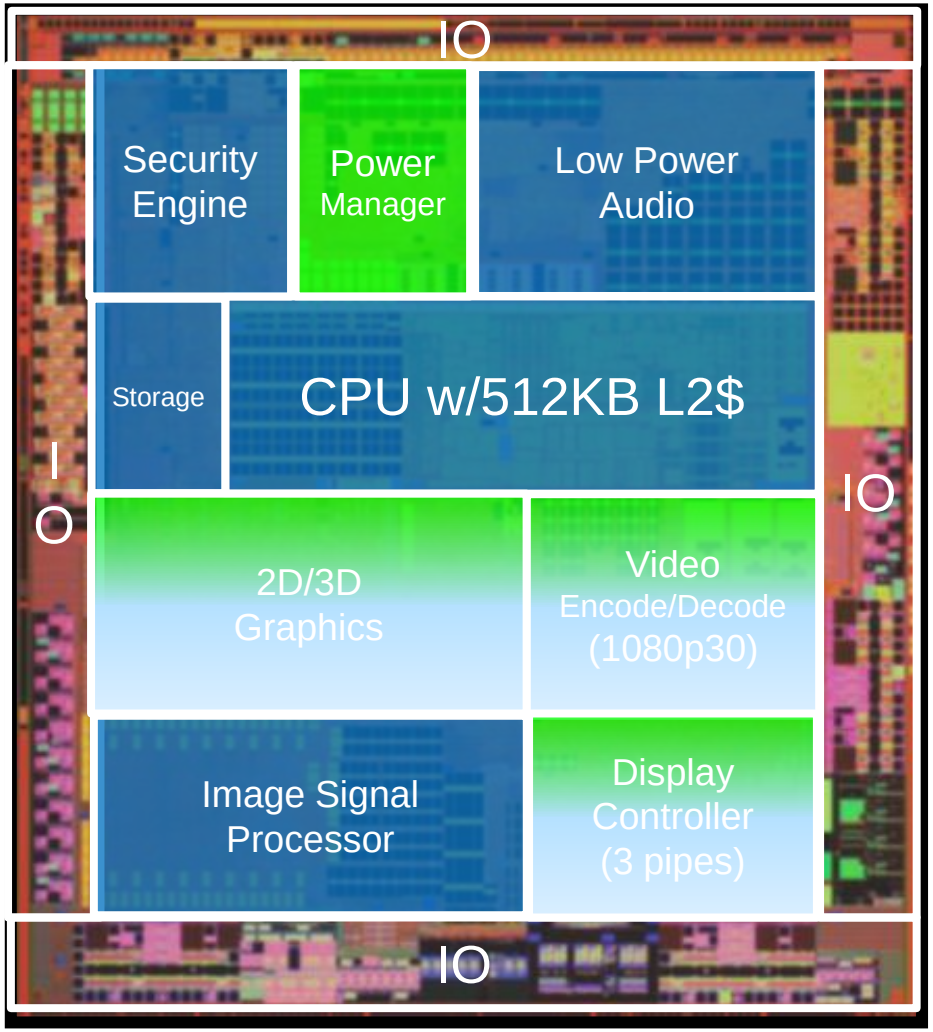
CPU Off

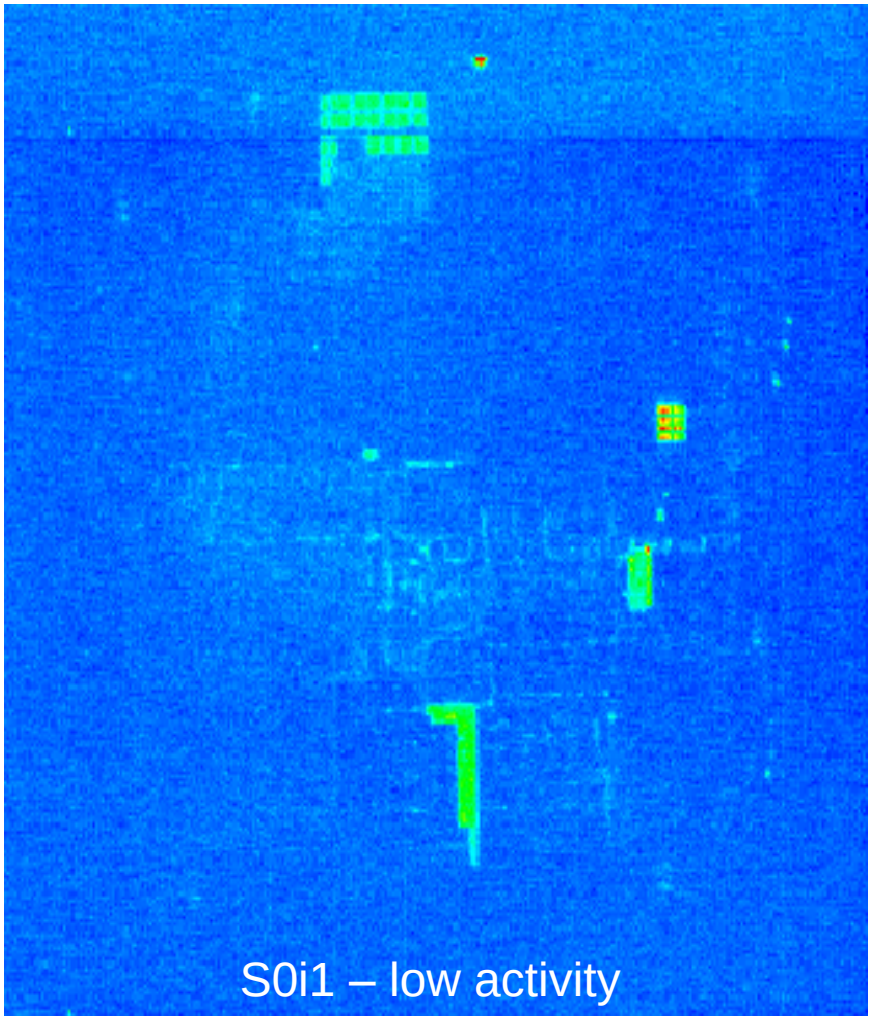




intel

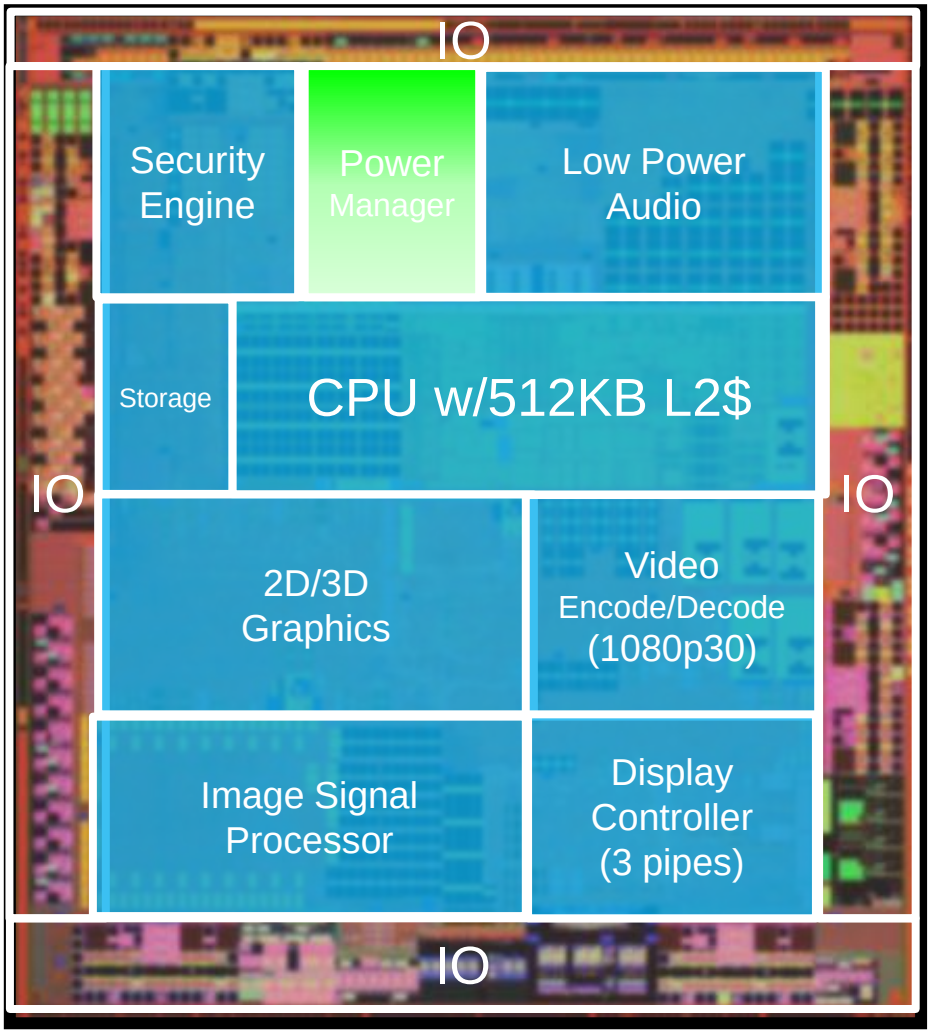
S0i1 System State

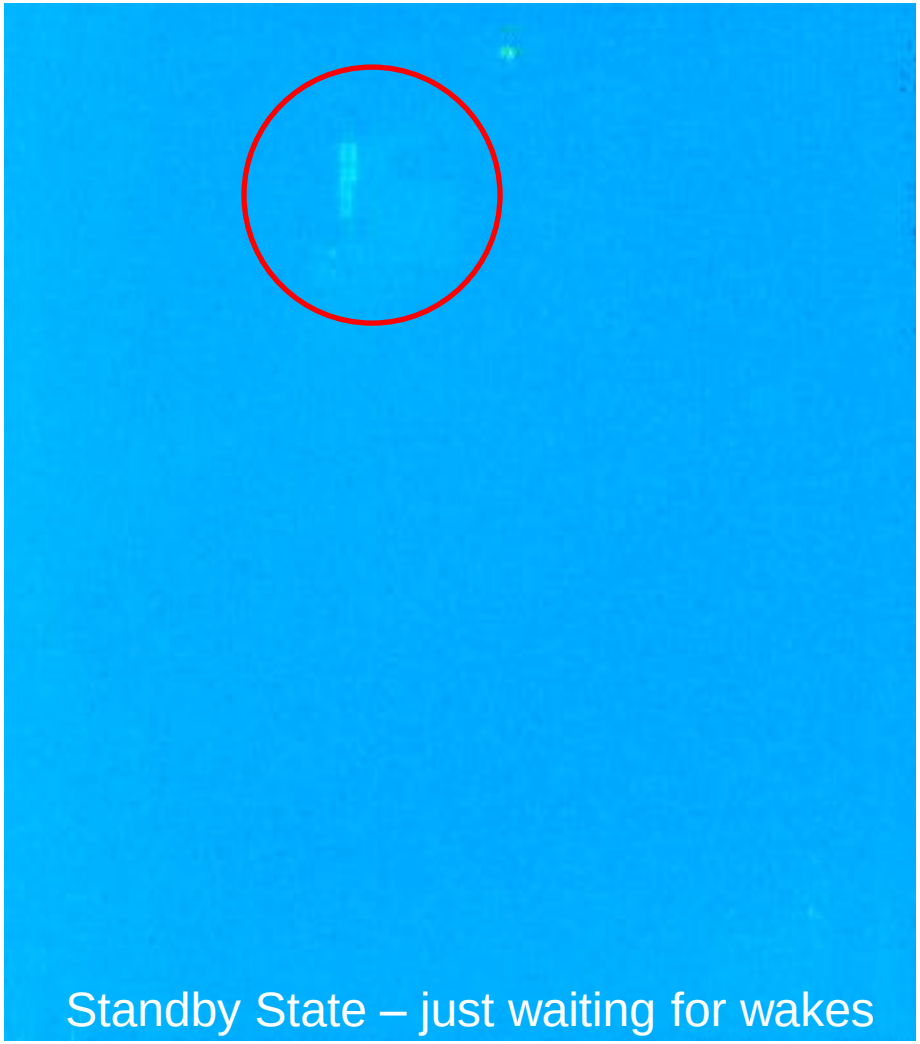




intel

S0i3/S3 System State

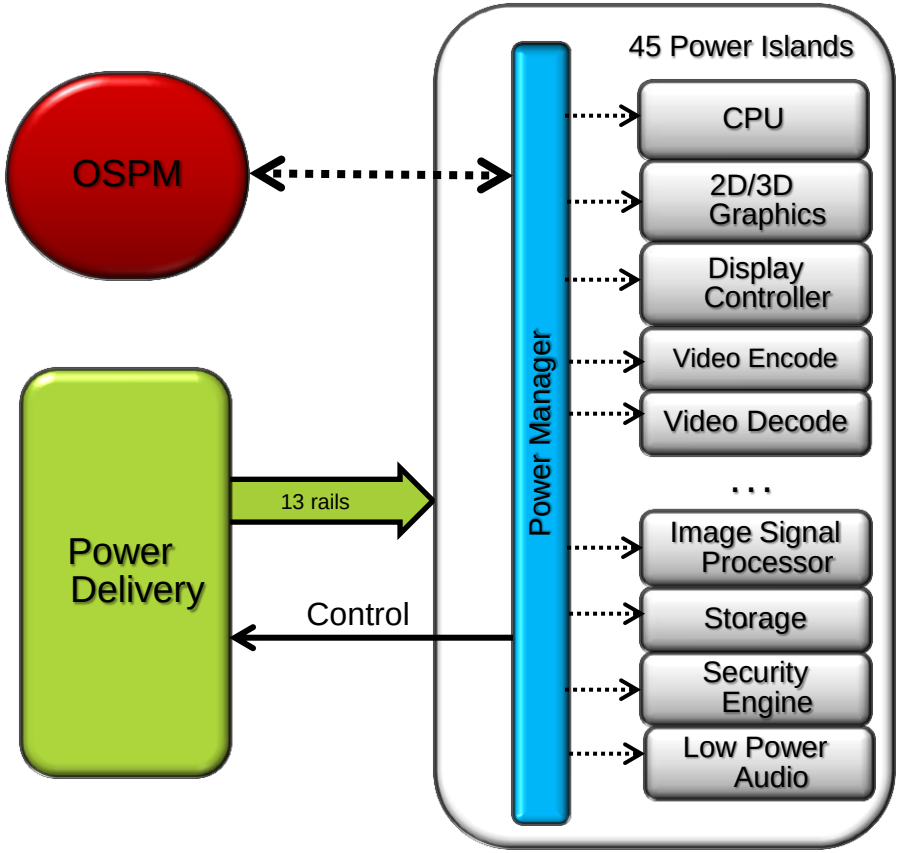




intel

New OS Power Management (OSPM)

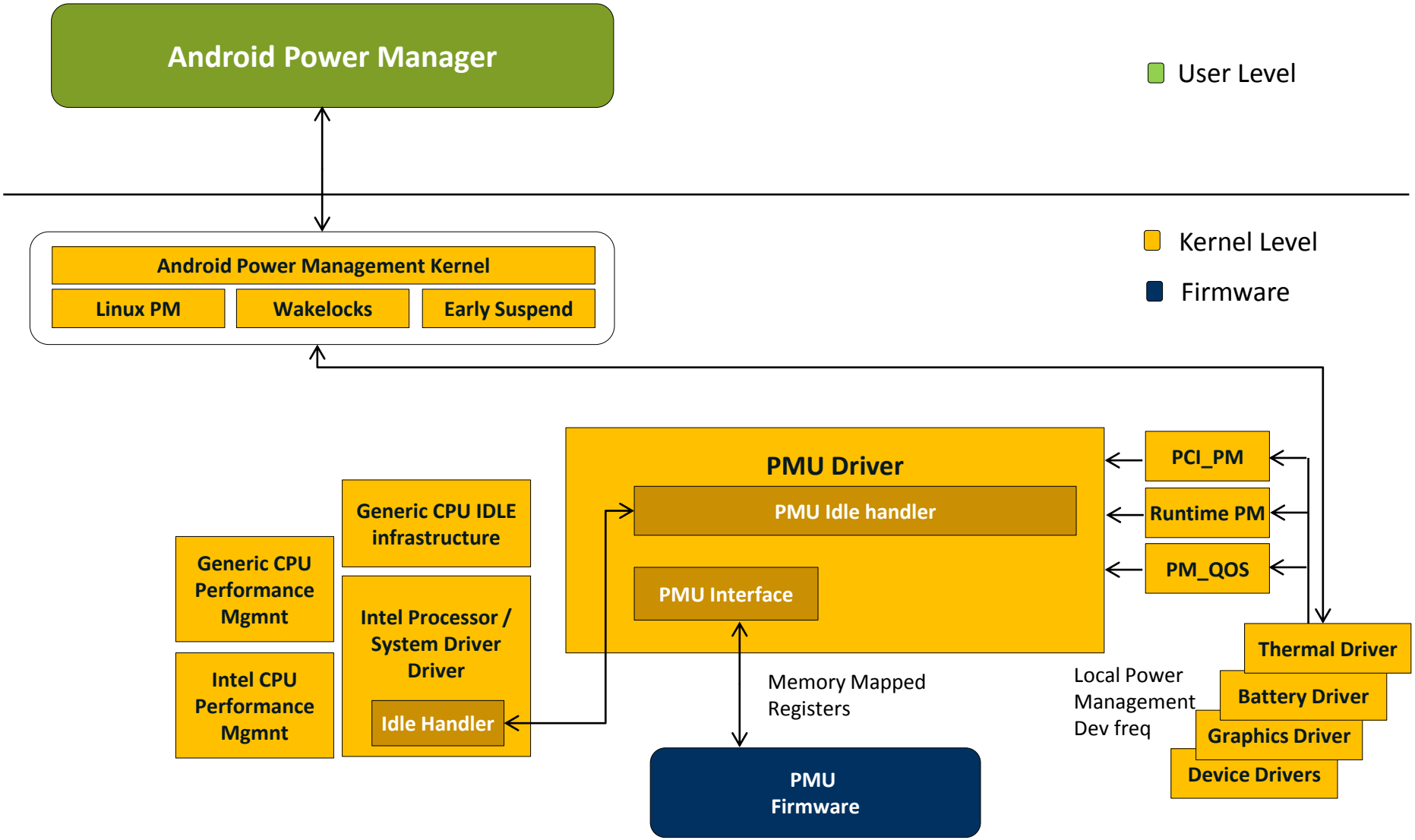
- Pervasive Power Management
 - ✓ Integrated PMU
 - ✓ Dedicated Power Delivery IC
 - ✓ Active management through HW, FW, SW
- Software-Directed
 - ✓ Operating system power management
 - ✓ Manages all hardware capabilities
- Fine Grain Power Management
 - ✓ 13 rails for IO & logic voltages
 - ✓ 45 Power islands for sub-systems
 - ✓ Aggressive power and clock gating
 - ✓ Integrated clocks and VR power down



OSPM Directs Entire Platform to Lowest Power State

intel

Platform Power Management Architecture



Medfield Reference Platform

High Performance CPU
1.6 Ghz Intel® Atom™ Processor Z2460

Full HD Video
1080p, 30fps Video Encoding
1080p, 30fps Video Playback

Advanced Imaging
Intel Image Signal Processing (ISP)
Advanced UI/UX from Intel

Great Graphics
PowerVR SGX 540 @ 400 MHz

High Speed Connectivity
Intel XMM 6260 21/5.8Mbps HSPA+

Apps
Google* Play (Android* Apps)



High Resolution Display
Internal : 1024x600;1024x768p capable
External : 1920x1080 p30 & i60

Optimized Android Support
Customizable User Experience

Enhanced Power/Batterylife
Standby: 14 days**
Video (1080p): 6 hours
Browsing 3G: 5 hours
Voice Call: 8 hours

Security
Programmable Security Engine
Remote Management Features

Operating System
Android 2.3.7 (Gingerbread)
Android 4.0.4 (Ice Cream Sandwich)

Current Design Wins
Lava XOLO X900 in India
Lenovo K900 in China
Orange San Diego in UK
Orange with Intel Inside® in France.

* other brands and names may be claimed as the property of others
** Battery: 1500mA, 3.7V

Smartphone Platform Roadmap



PERFORMANCE SMARTPHONE

Intel® Atom™ Z2460

Up to 1.6 GHz Processor
Intel XMM™ 6260: HSPA+

Higher Performance



Intel® Atom™ Z2580

Up to 2X Performance
2X HSPA+ / LTE

Lower Cost



VALUE SMARTPHONE

1.2 GHz Processor
Intel XMM 6265: HSPA+

All products, designs, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing.

Medfield Summary

- Medfield meets tight Smartphone power consumption constraints and provides outstanding scalar CPU performance
 - ✓ “Race to Idle” minimizes energy consumption while providing excellent end-user experience
 - ✓ Ultra low power SOC states cater to common “user idle” and “system idle” scenarios
 - ✓ Accelerators for Video, Camera, Audio processing provide energy optimized media capabilities





CLOUD
TRANSFORMS IT

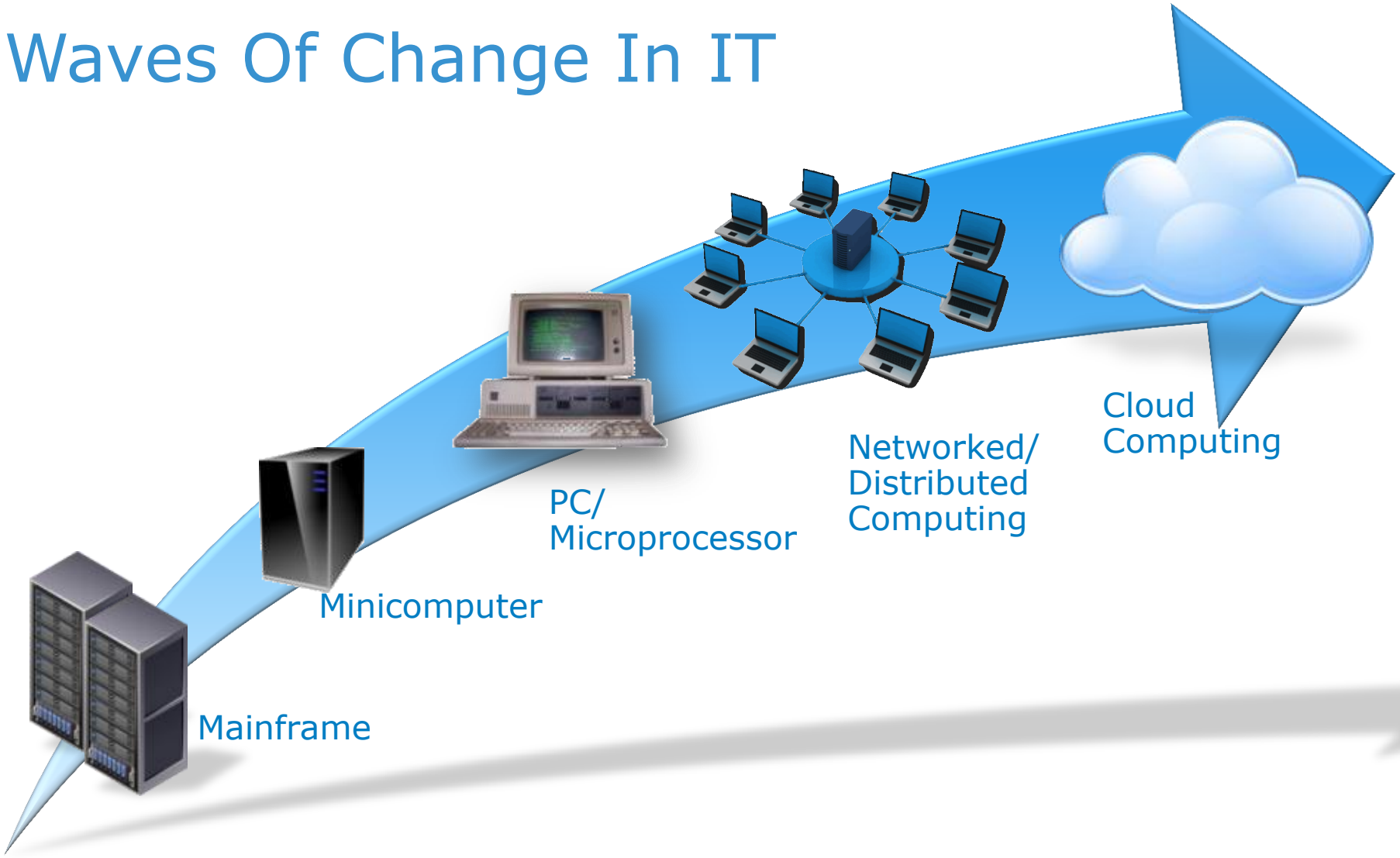
BIG DATA
TRANSFORMS BUSINESS

Pat Gelsinger
President & COO
Information Infrastructure Products
EMC Corporation

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.

Waves Of Change In IT



© Copyright 2012 EMC Corporation. All rights reserved.

CLOUD
TRANSFORMS IT



© Copyright 2012 EMC Corporation. All rights reserved.

Phases Of IT Maturity



Packaged Applications
Flat IT Tax, Project-centric
Dedicated Vertical Stacks

Reactive
Respond To Business Request



Service Catalog
Cost & Use Metrics
Dynamic Resource Pools

Proactive
Increase IT Agility



New Business Applications
Pay-For-Use
Automated Infrastructure

Innovative
Differentiate The Business

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.

Steps Of IT Transformation



Cloud Infrastructure
Build A Software-Defined Data Center

Cloud Operating Model
Standardize & Automate To Run IT-As-A-Service

Cloud Applications
Build Mobile & Predictive Apps



Reactive
Respond To Business Request

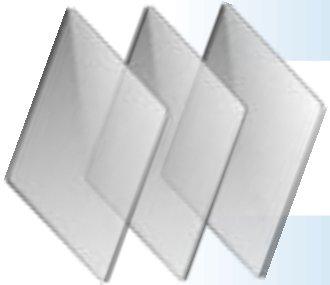
Proactive
Increase IT Agility

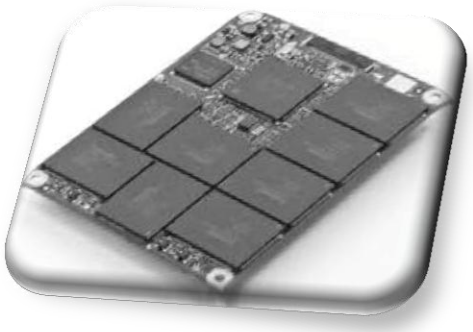
Innovative
Differentiate The Business

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.





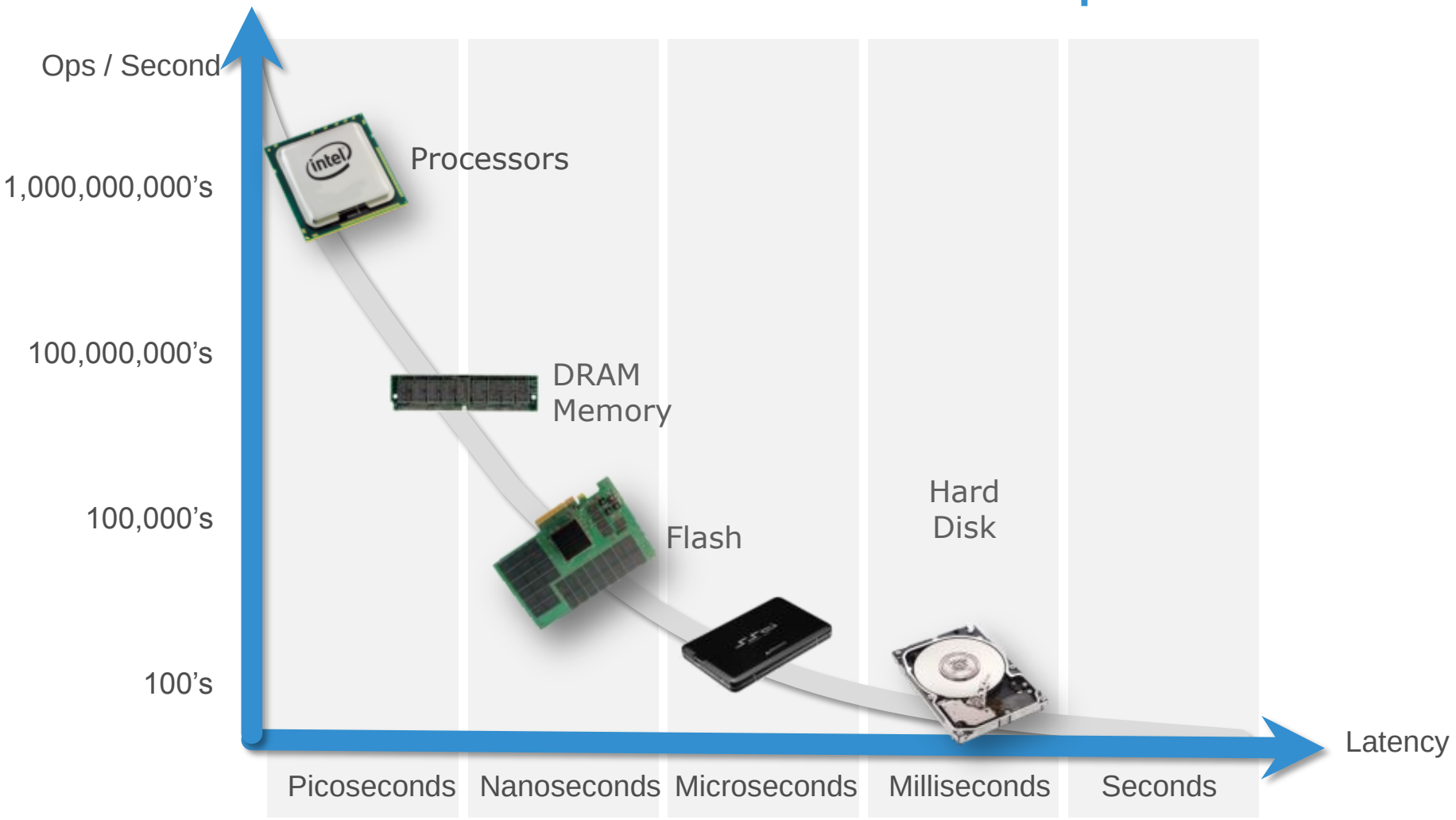




Cloud.

© Copyright 2012 EMC Corporation. All rights reserved.

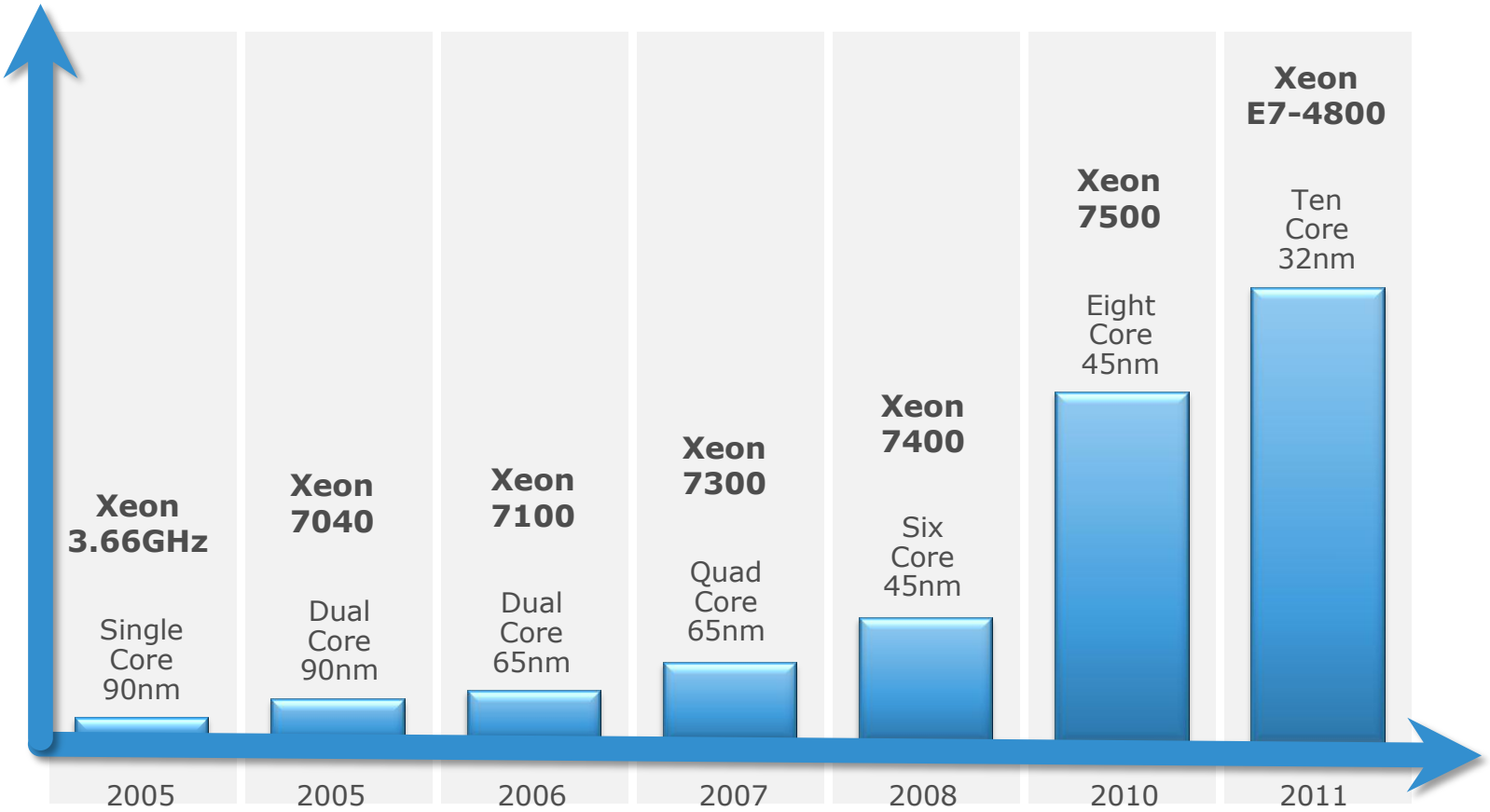
Flash Fills The Performance Gap



© Copyright 2012 EMC Corporation. All rights reserved.

Dramatic Performance Growth For x86

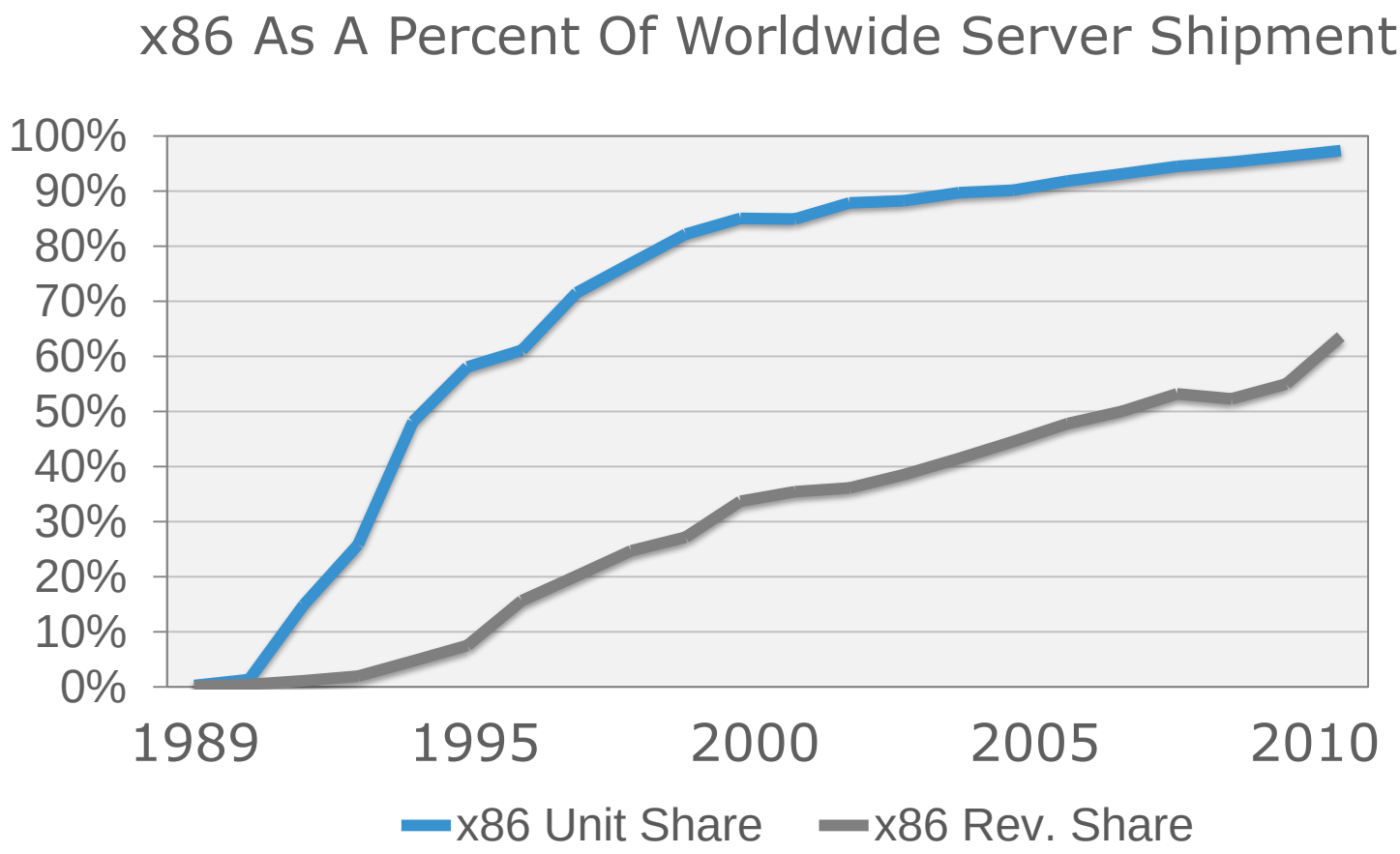
2000% Performance Increase Since 2005



Source: Intel internal OLTP database workload performance estimates as of 15 April 2011. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

© Copyright 2012 EMC Corporation. All rights reserved.

Dominant Market Share For x86

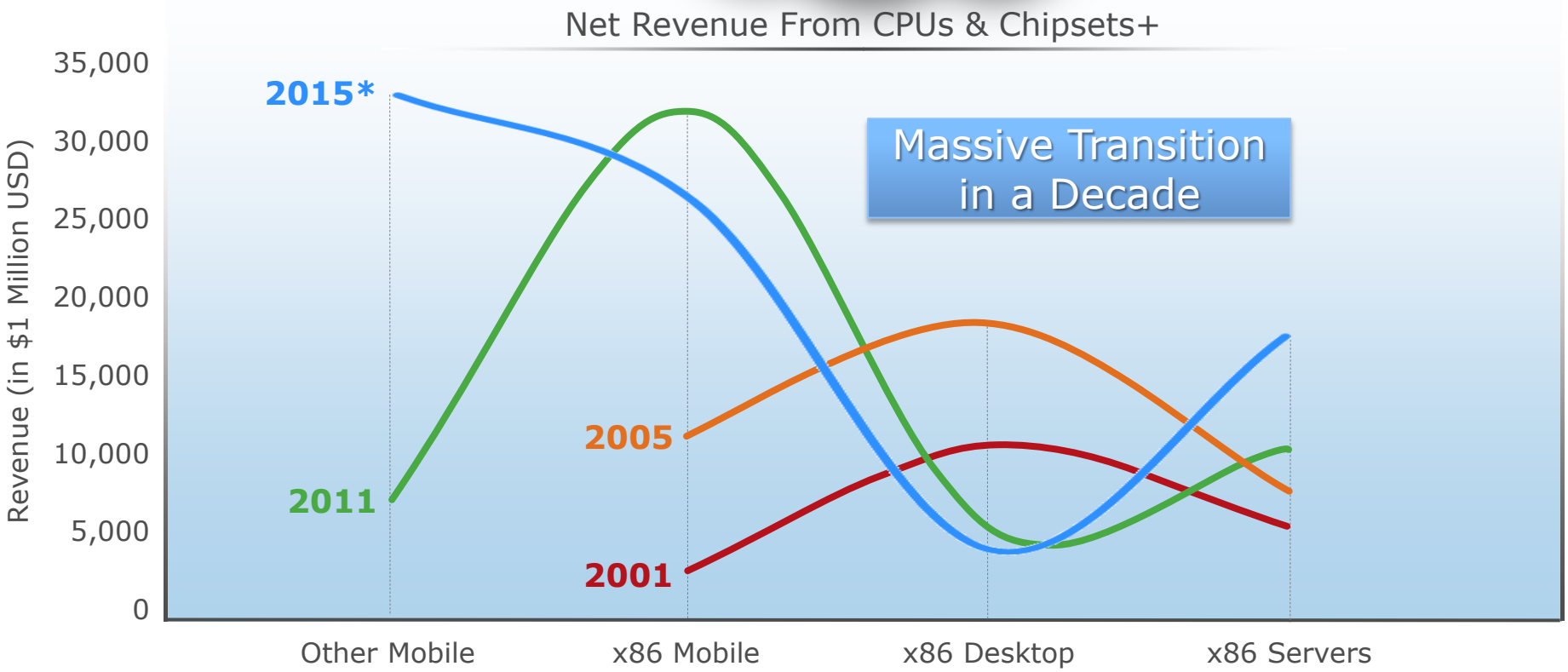


Source: IDC

© Copyright 2012 EMC Corporation. All rights reserved.

Waves of Change





+: Data from Intel 2011, 2005, 2001 Annual Reports, http:// www.isuppli.com, http:// www2.uta.edu/marketing

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.

Future Si Design - Mobile

More than Power, Performance, Cost and Footprint

- Embed HW in SoC for:
 - Virtualization
 - Graphics: Remote desktop / graphics-rich remote UI
 - Security: e-Currency, DRM, Anti-Virus
 - Encryption: Fast, secure end-point communication
- Standard HW interface for generic OS / SW management



Cloud.

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.

Future Si Design- Server

More than Power, Performance, Cost and Footprint

- Design for the Software-Defined Data Center & Big Data
 - Server NICs integrate VXLAN VTEP
 - HW accelerates remote graphics-rich desktops and connection protocols
 - Programmable HW to classify and inspect network packets
 - Large on-chip, high-speed memory (SRAM, PCM, Flash)

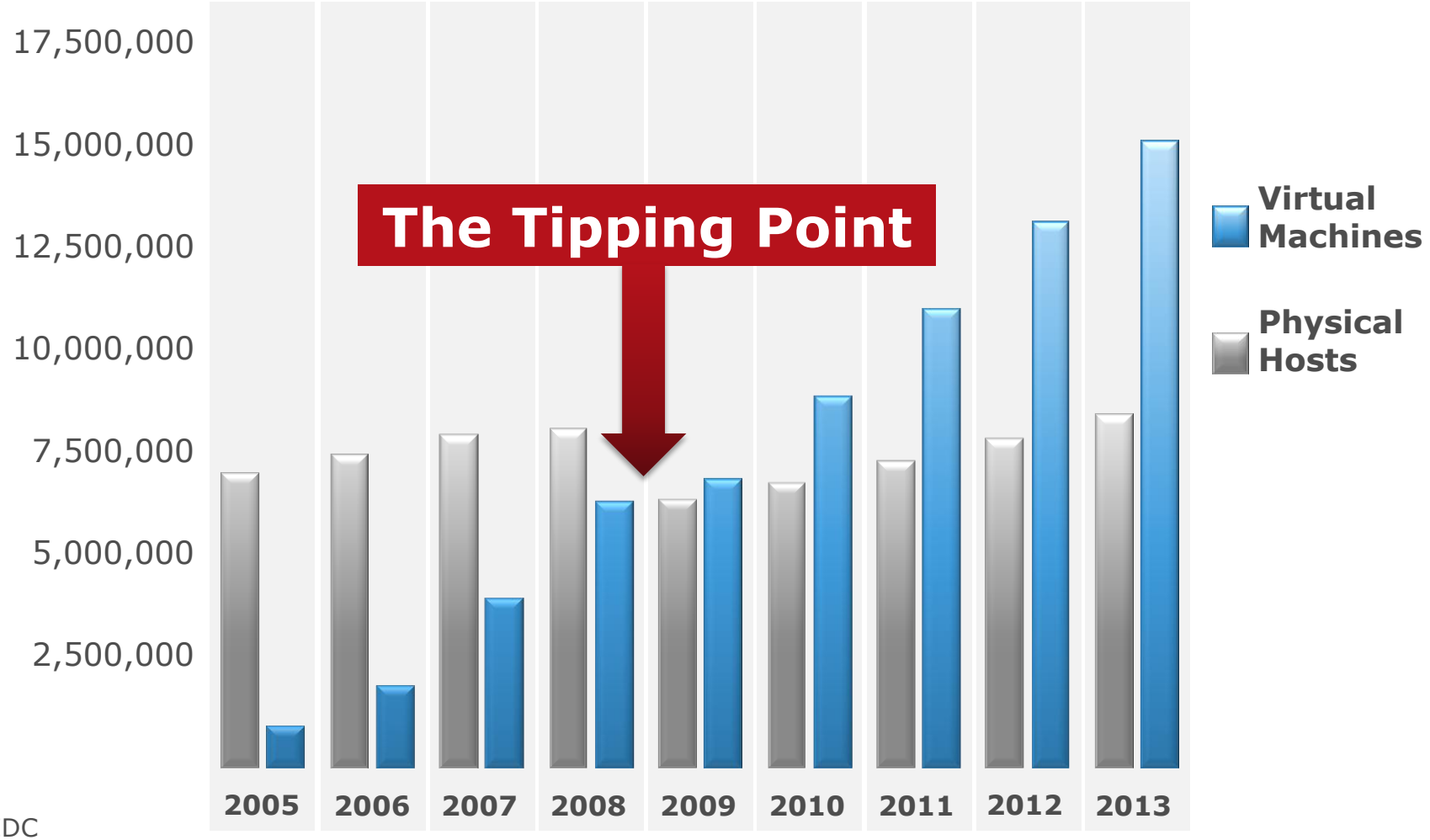


Cloud.

EMC²

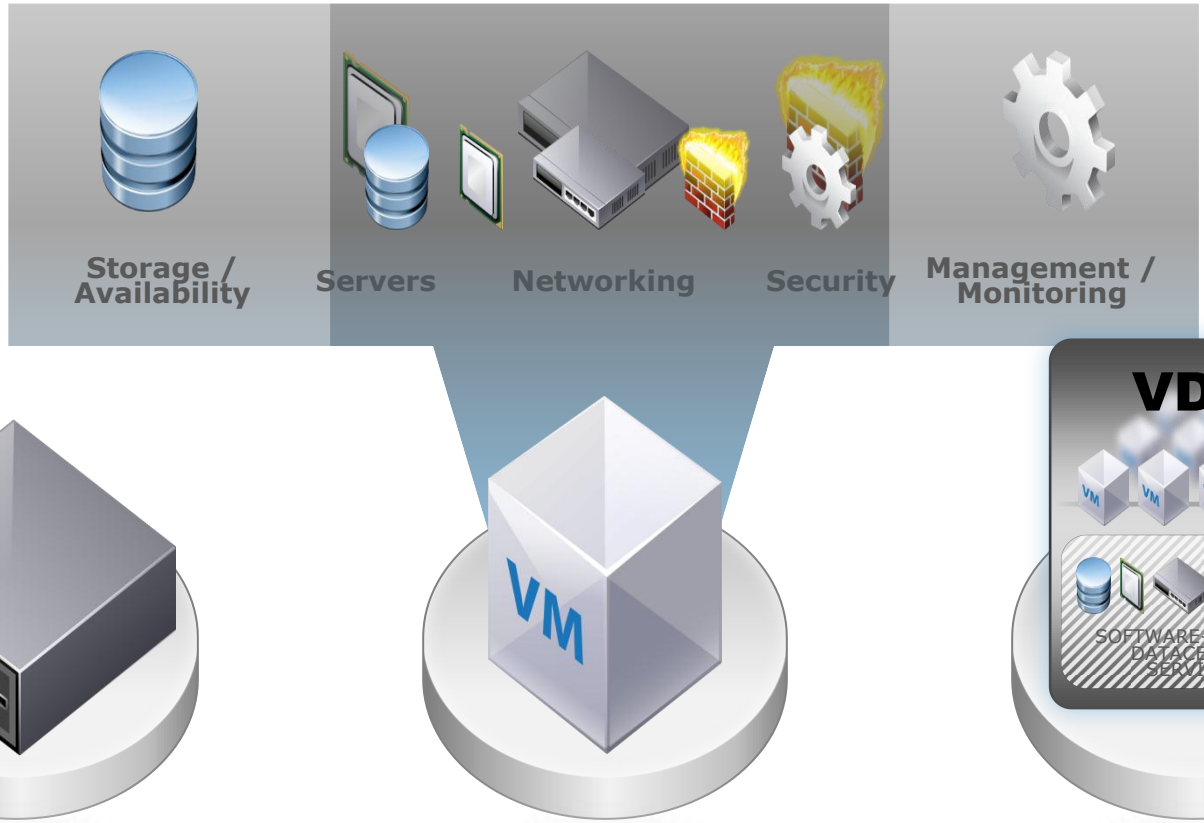
© Copyright 2012 EMC Corporation. All rights reserved.

2009: More Apps On Virtual Infrastructure



Source: IDC

© Copyright 2012 EMC Corporation. All rights reserved.



WEEKS

2008

DAYS / HOURS

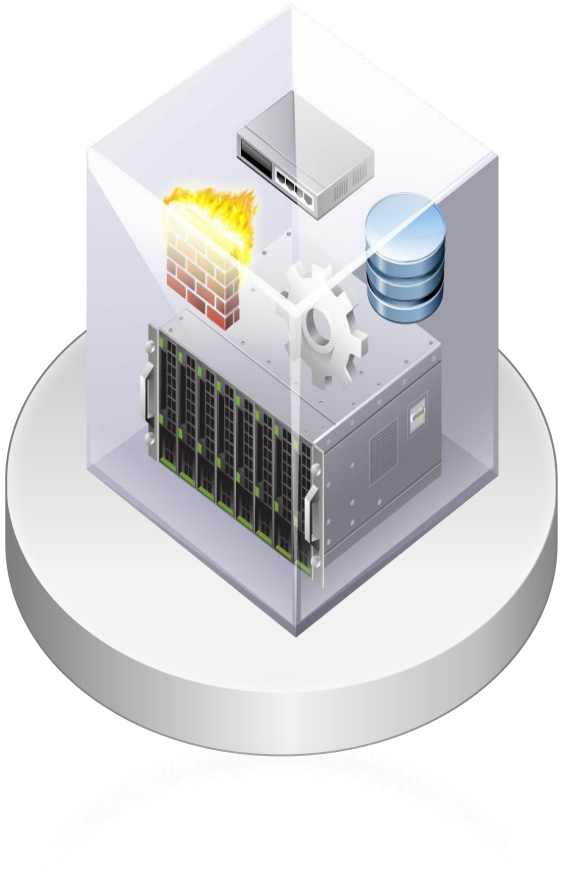
2012

MINUTES / SECONDS

FUTURE

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.

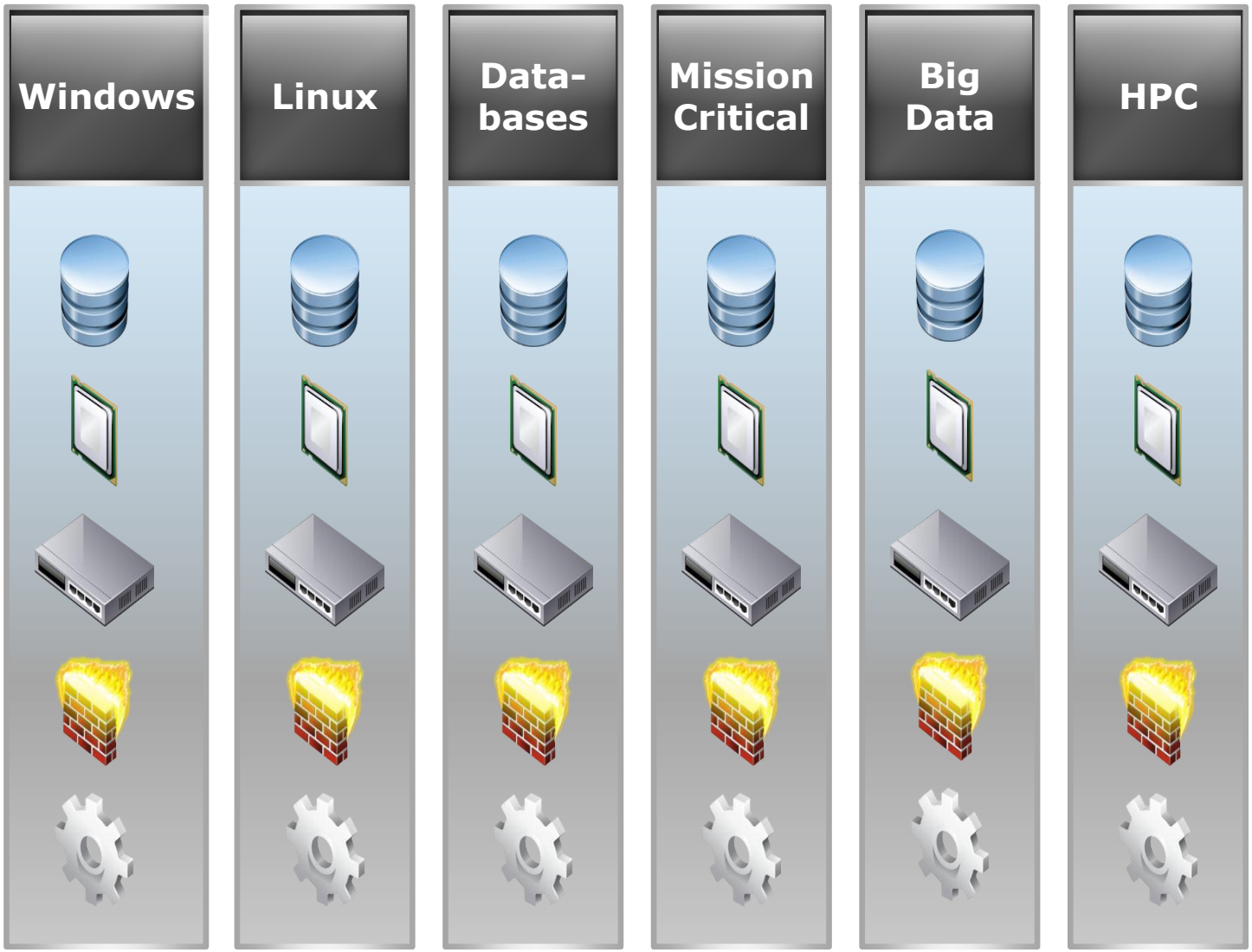


SOFTWARE-DEFINED DATACENTER

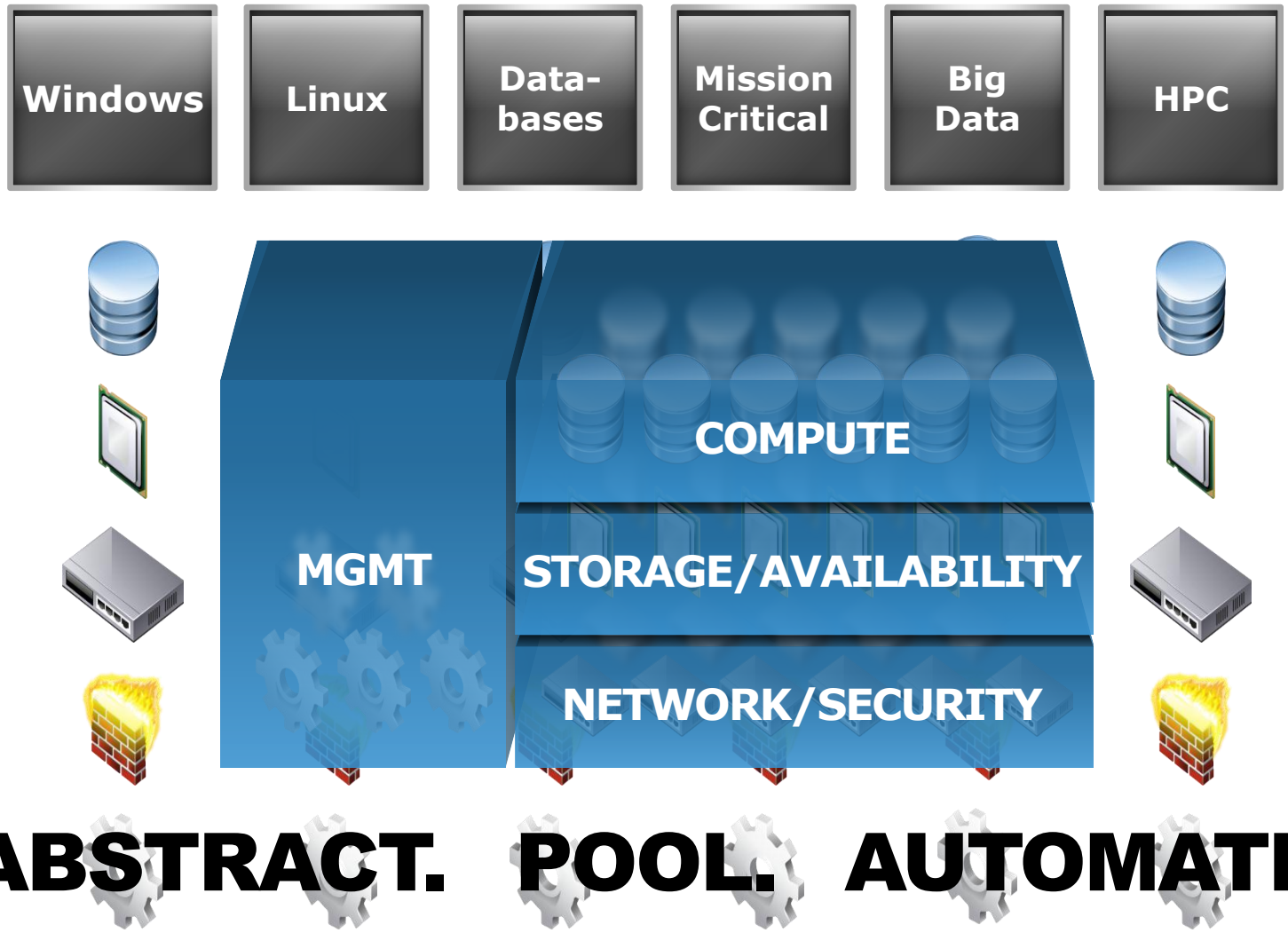
All infrastructure is virtualized and delivered as a service, and the control of this datacenter is entirely automated by software.

© Copyright 2012 EMC Corporation. All rights reserved.

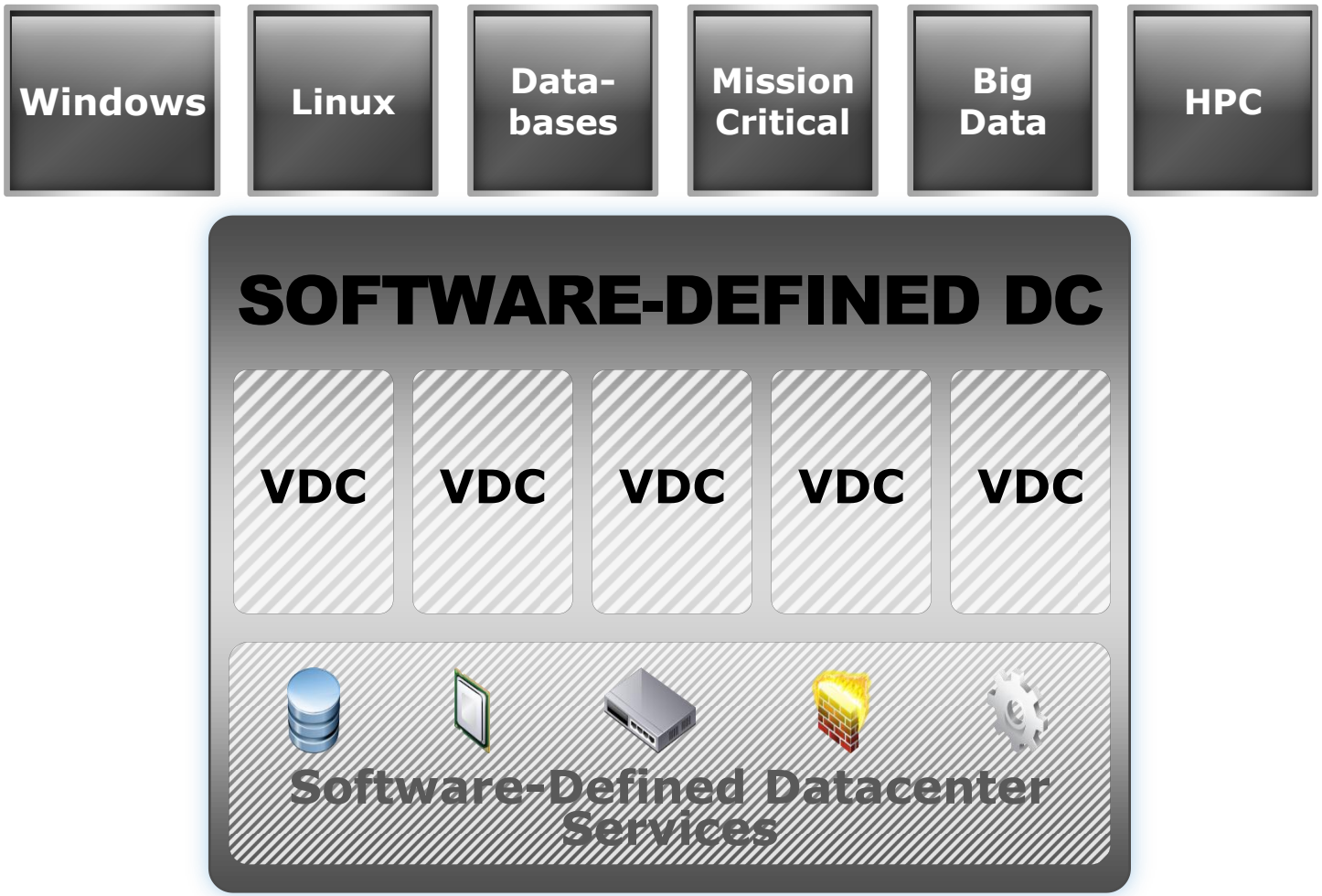
Traditional View of the DC Environment



© Copyright 2012 EMC Corporation. All rights reserved.



© Copyright 2012 EMC Corporation. All rights reserved.



© Copyright 2012 EMC Corporation. All rights reserved.

TRUST TRANSFORMS CLOUD



© Copyright 2012 EMC Corporation. All rights reserved.

Old World: Static Security



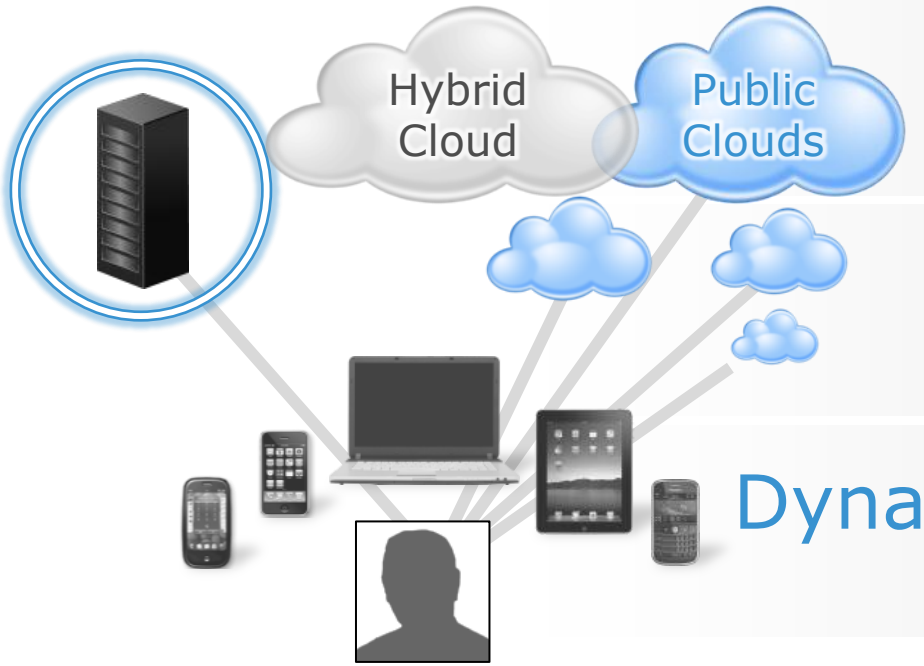
Static Attacks
Generic, Systems-Based

Static Infrastructure
Physical, IT Controlled

Static (Bolt-On) Defenses
Signature-Based, At Perimeter

© Copyright 2012 EMC Corporation. All rights reserved.

New World: Dynamic Security



Dynamic Attacks
Targeted, Human-Based

Dynamic Infrastructure
Virtual, User-Centric

Dynamic (Built-In) Defenses
Analytics & Risk-Based

© Copyright 2012 EMC Corporation. All rights reserved.



BIG DATA TRANSFORMS BUSINESS

© Copyright 2012 EMC Corporation. All rights reserved.



Sources: "How Much Information?" Peter Lyman and Hal Varian, UC Berkeley,. 2011 IDC Digital Universe Study.

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.



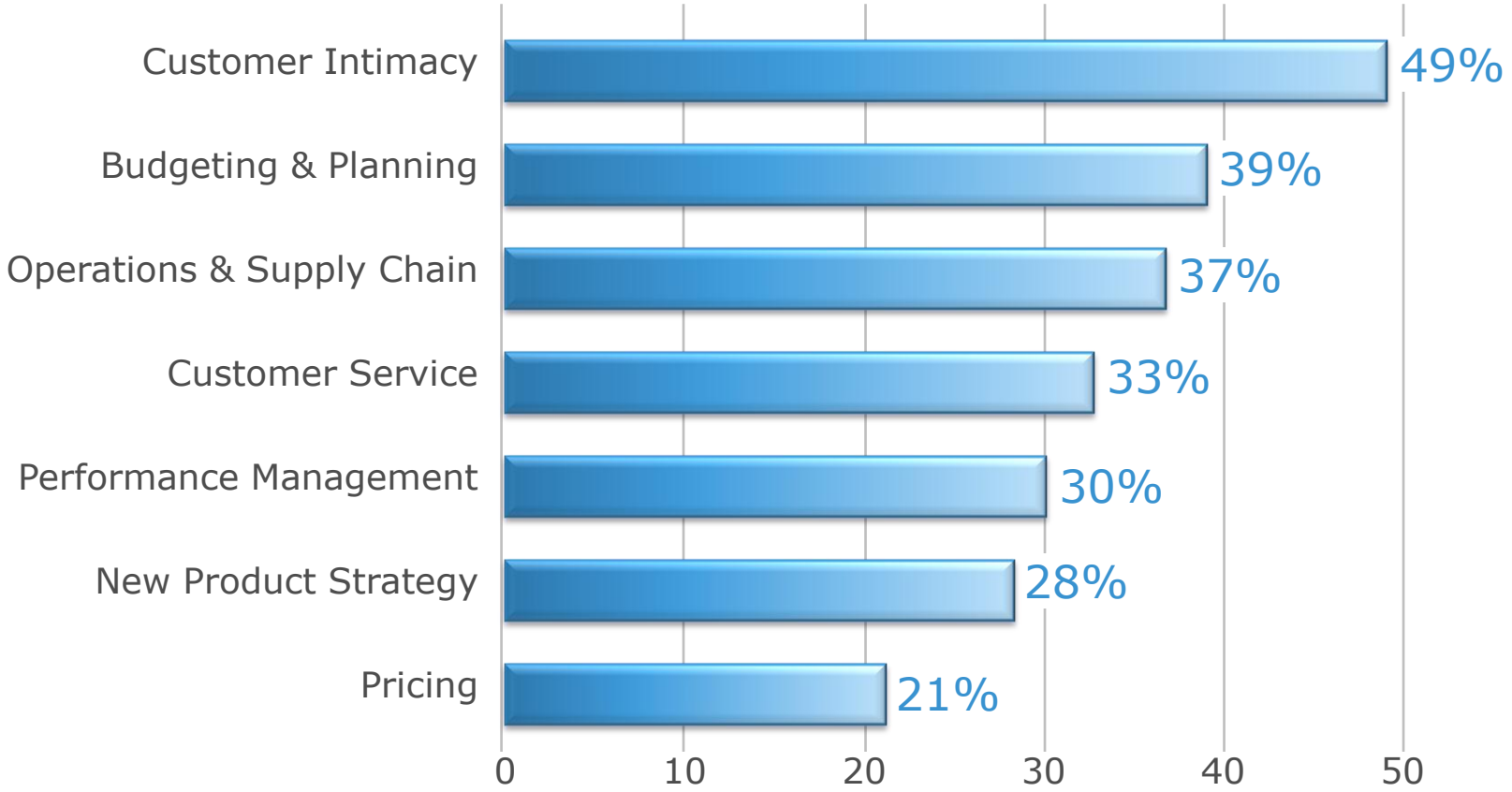
Sources: "How Much Information?" Peter Lyman and Hal Varian, UC Berkeley,. 2011 IDC Digital Universe Study.

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.

How Companies Are Using Big Data

Functional Areas Where Companies Are Using Big Data



McKinsey Global Survey of 1,469 C-level executive respondents at a range of industries and company sizes, "Minding Your Digital Business," 2012.

© Copyright 2012 EMC Corporation. All rights reserved.

big•data \ datasets so large they break traditional IT infrastructures.



EMC²

© Copyright 2012 EMC Corporation. All rights reserved.

BI focuses on managing and reporting on **existing data** to **monitor** and **manage** concerns within the enterprise



Data Science applies advanced **analytical** tools and algorithms to generate **predictive insights** and **new** product **innovations** that are a direct result of the data

Who Is The Data Scientist?

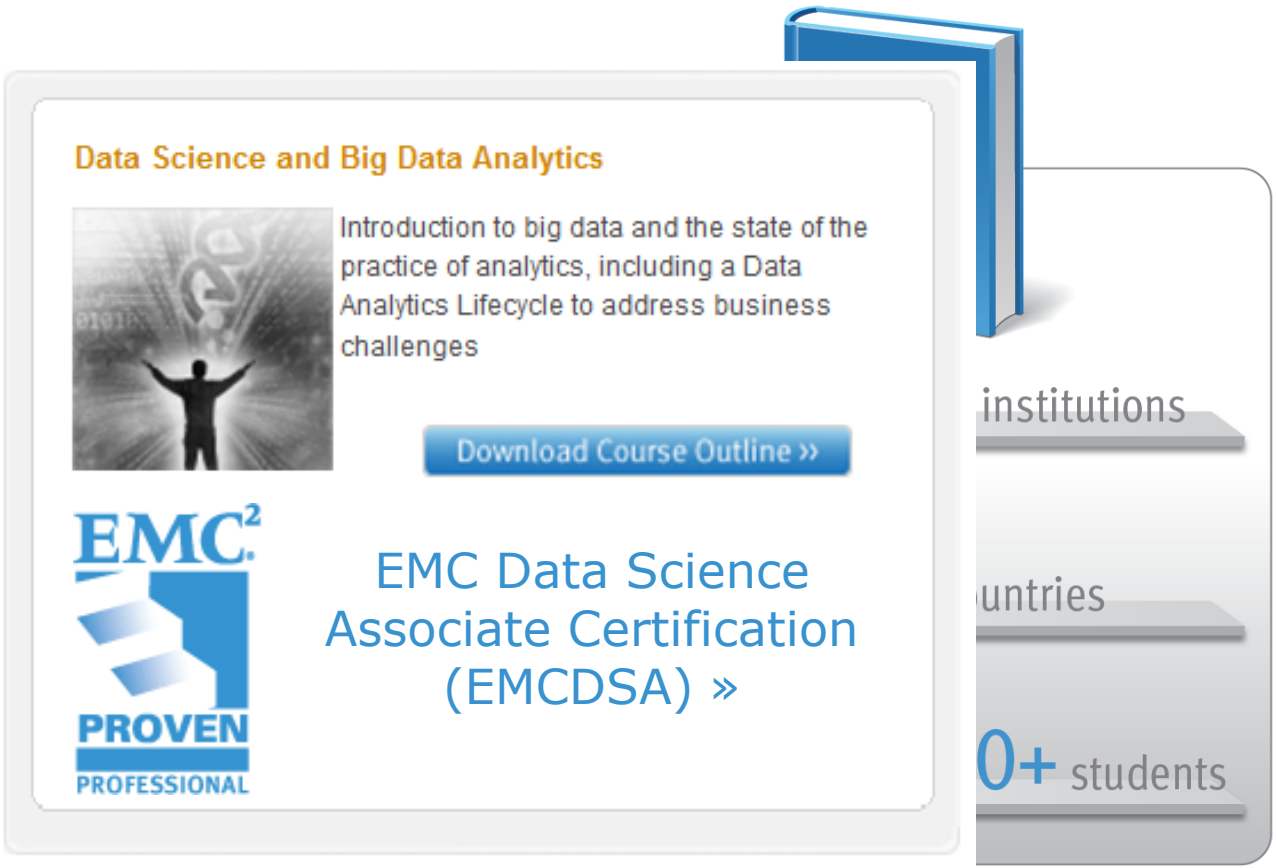
Source: EMC Study, "Data Science Revealed: A Data-Driven Glimpse into the Burgeoning New Field," December 5, 2011

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.

Training Tomorrow's Talent

EMC Academic Alliance



© Copyright 2012 EMC Corporation. All rights reserved.



In Summary

- Silicon design remains essential – HW/SW co-design is critical
- The action is in the edges (Mobile & Server)
- Cloud becomes the Software-Defined Datacenter
- Big Data opens up new opportunities for HW design

EMC²

© Copyright 2012 EMC Corporation. All rights reserved.

EMC²

POWER7+™

IBM



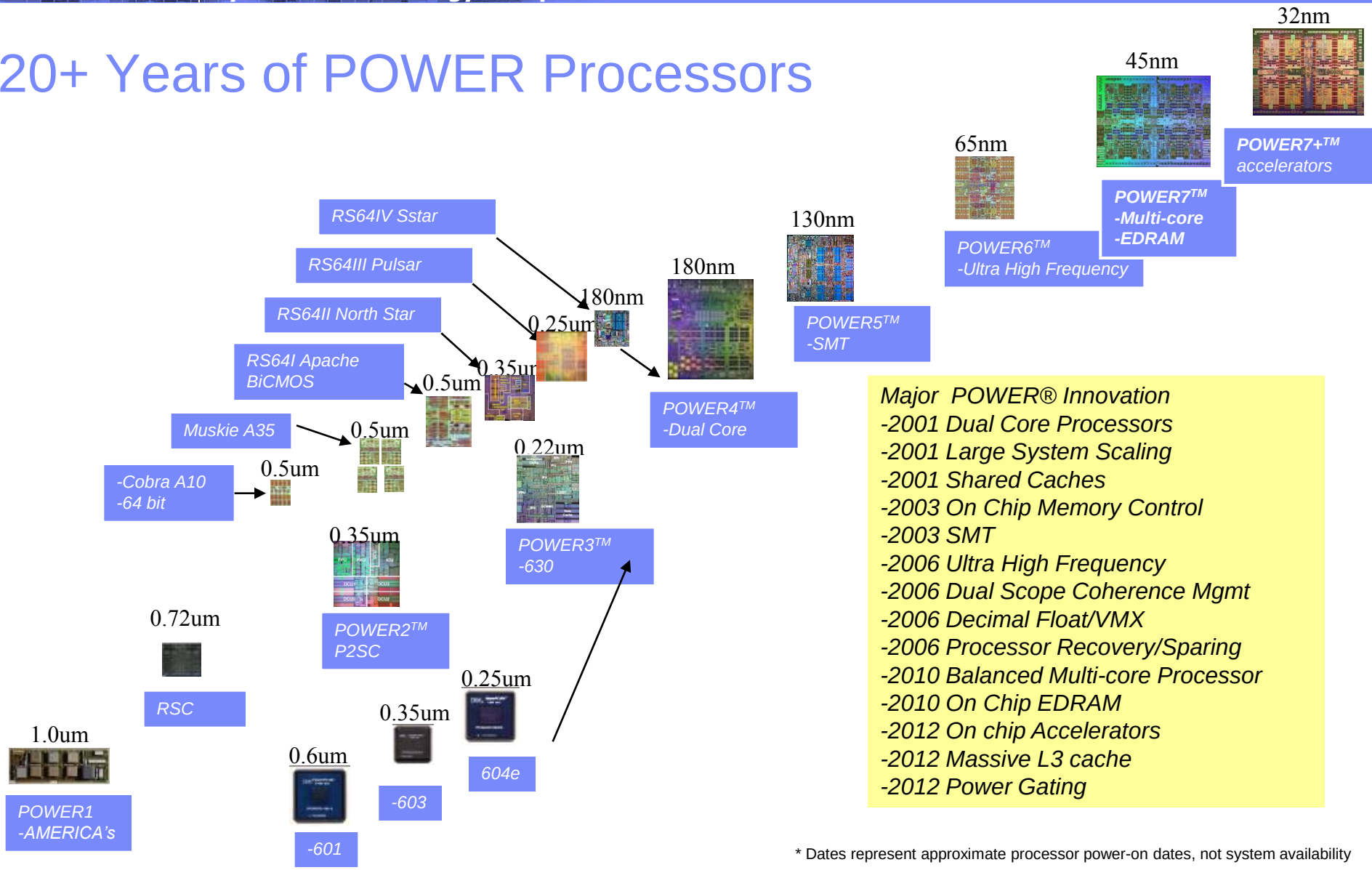
© 2012 IBM Corporation

Outline

- POWER Processor History
- Design Overview
- Performance Benchmarks
- Key Features
 - Scale-up / Scale-out
 - The new accelerators
 - Advanced energy management
- Summary

* Statements regarding Power7+™ features do not imply that IBM will introduce a system with this capability

20+ Years of POWER Processors

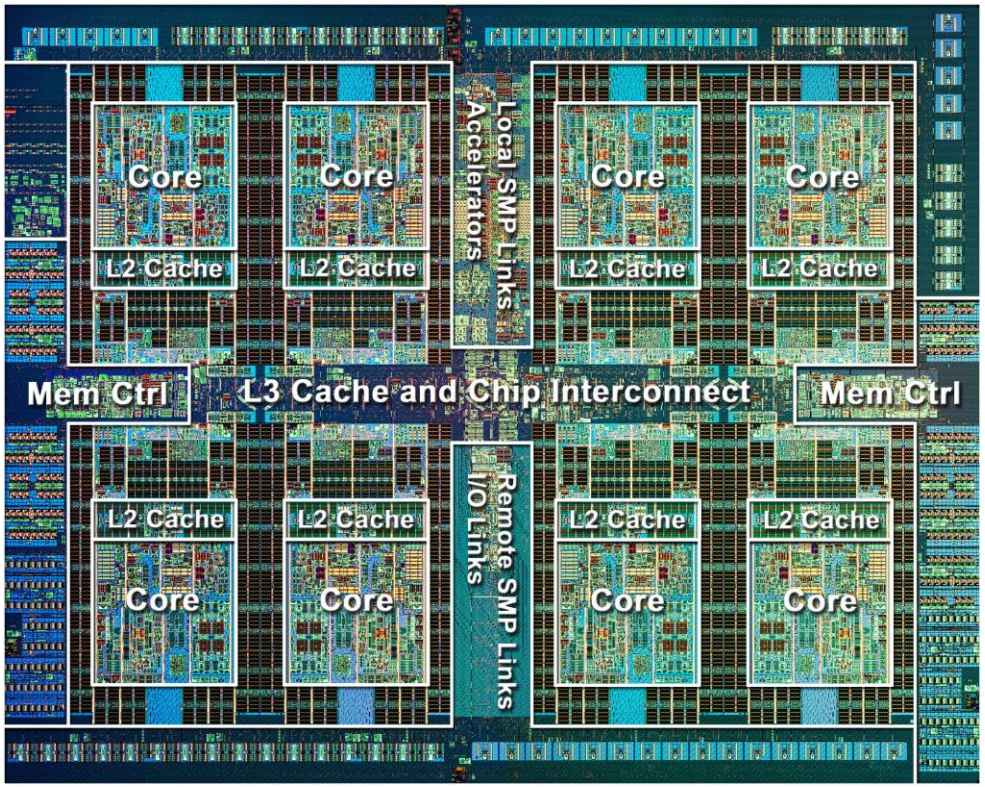


Major POWER® Innovation
-2001 Dual Core Processors
-2001 Large System Scaling
-2001 Shared Caches
-2003 On Chip Memory Control
-2003 SMT
-2006 Ultra High Frequency
-2006 Dual Scope Coherence Mgmt
-2006 Decimal Float/VMX
-2006 Processor Recovery/Sparing
-2010 Balanced Multi-core Processor
-2010 On Chip EDRAM
-2012 On chip Accelerators
-2012 Massive L3 cache
-2012 Power Gating

* Dates represent approximate processor power-on dates, not system availability

POWER7+ Processor Chip

- Area: 567mm2
- Eight processor cores
 - 12 execution units per core
 - 4 Way SMT per core
 - 32 Threads per chip
 - 256KB L2 per core
- Scalability up to 32 Sockets
 - 360GB/s SMP bandwidth/chip
 - 20,000 coherent operations in flight
- Technology: 32nm lithography, Cu, SOI, eDRAM, 13 metal levels
- 2.1B transistors
 - Equivalent function of 5.4B
- 80MB on chip eDRAM shared L3
- Accelerators
- Enhanced Power management
- Binary Compatibility with POWER6/7

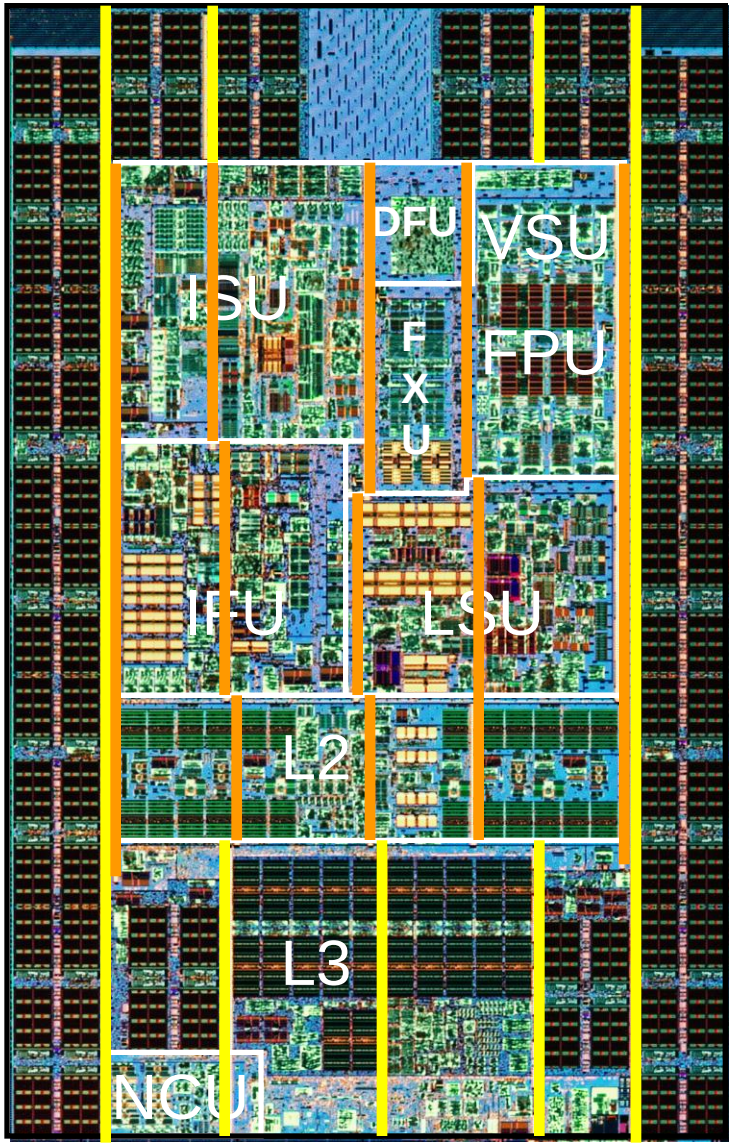


An Improved Core

- Up to 25% frequency gain due to mapping into 32nm technology and power management improvements.
- Increase of L3 memory capacity by 2.5x
- Doubled single precision floating-point performance
- Added Power Gating regions for Core/L2 & L3 regions

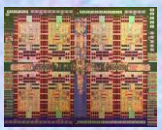
Core/L2 Power-Gating

L3 Power-Gating



Optimized in Two Dimensions

POWER7+



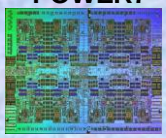
Single Chip Module

Increased Clock Speed → Thread Strength

2.5x L3 cache → Thread Strength
SMP Scaling

Scale UP

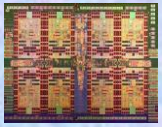
POWER7

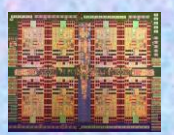


Single Chip Module

Scale OUT

POWER7+



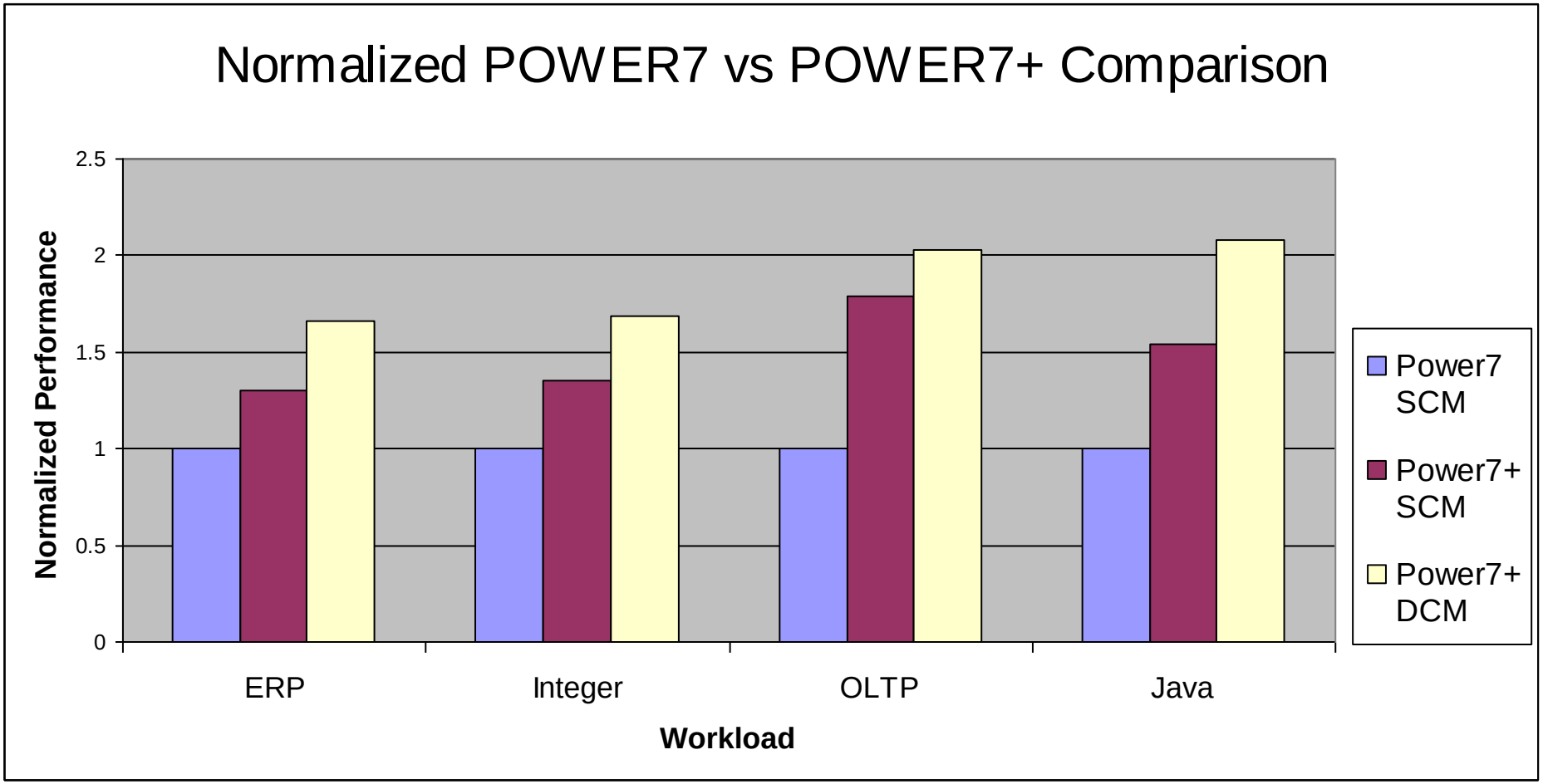


Dual Chip Module

2x Cores / socket → Throughput

5x L3 cache → Bandwidth Amplifier

Performance View

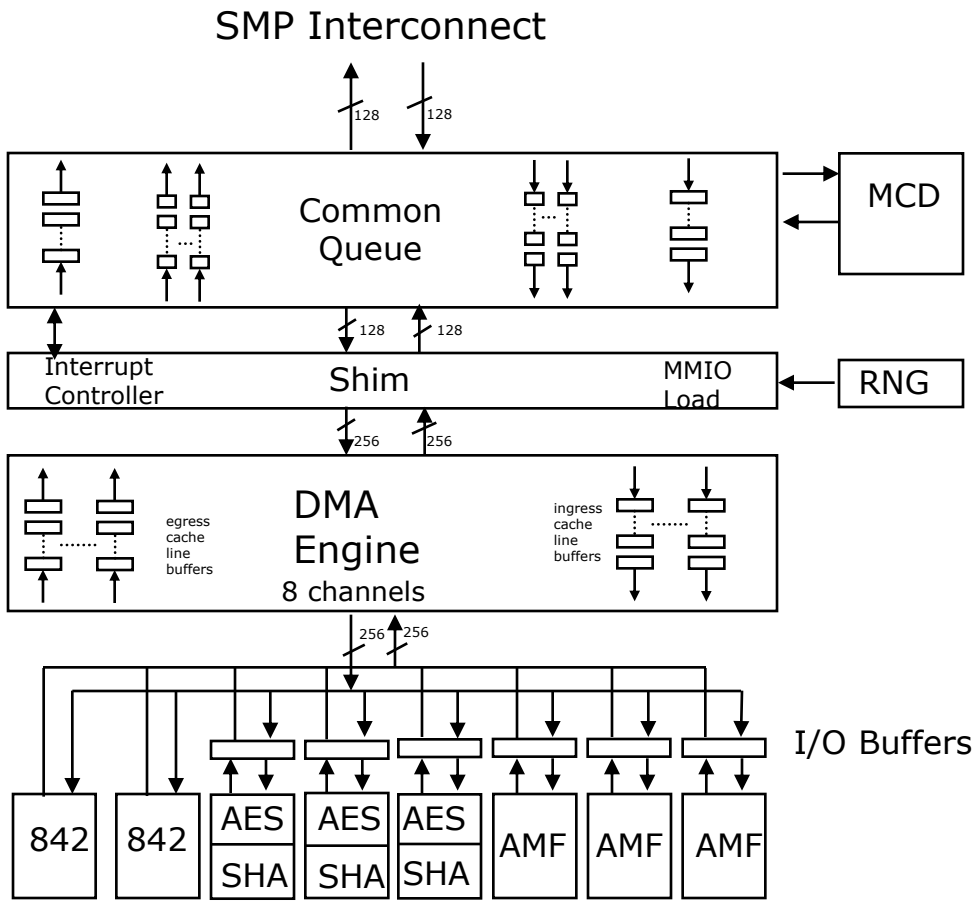


POWER7+ Accelerators

- Provide CPU off-load and workload speedup for SSL, encrypted file system, and active memory expansion (AME).
 - Asymmetric Math Functions (AMF)
 - RSA cryptography
 - ECC (elliptic curve cryptography)
 - Advanced Encryption Standard (AES)/Secure Hash Algorithm (SHA)
 - Symmetric-key cryptography with combinational modes
 - Random Number Generator (RNG) – True hardware entropy generator
 - Cannot be algorithmically reverse engineered
 - 842 proprietary compression algorithm
 - High bandwidth, area efficient
- Integrated across silicon, ISA, hypervisor, and OS

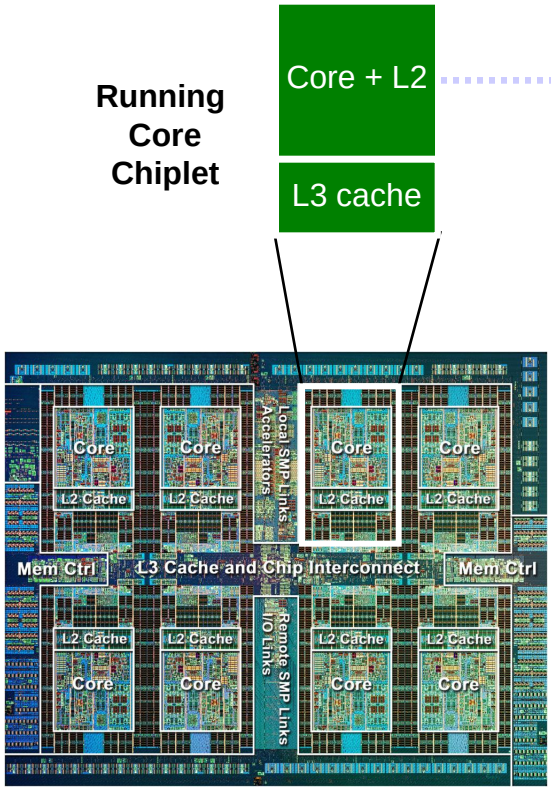
Accelerators Details

- Advanced Encryption Standard engine
 - Modes: ECB, CBC, CTR, CCM, CCA, GCM, GCA, GMAC, CM, F8, XBC-MAC-96
 - Key lengths: 128b,192b, 256b
 - Three engines
- Secure Hash Algorithm engine
 - Modes: SHA-1, SHA-256, SHA-512, MD5
 - HMAC supported for SHA
 - Three engines
- Asymmetric Math Functions
 - Modular math functions for RSA (Rivest, Shamir, Adleman) and ECC (elliptic curve cryptography): mod add, mod subtract, mod inverse, mod reduction, mod multiplication, mod exponentiation, mod exponentiation CRT(integer only)
 - Point functions for ECC GF(p) and GF(2m): point add, point double, point multiply
 - RSA lengths: 512b,1024b, 2048b, 4096b
 - ECC GF(p) lengths: 192b, 224b, 256b, 384b, 521b (SuiteB)
 - ECC GF(2m) lengths: 163b, 233b,283b, 409b, 571b (SuiteB)
- Random Number Generator
 - All digital design which produces 64b random numbers accessible by MMIO load instructions
 - Correctness verified against the NIST Random Number Generator Test Suite
- Active Memory Expansion
 - IBM-proprietary algorithm with 8B-, 4B-, and 2B-phrase parsings
 - Throughput: Up to 8 bytes of compression or 8 bytes of decompression per bus cycle.
- MCD
 - Hardware to predict whether memory access is on-node or off-node.



POWER7+ Sleep & Winkle Overview

Running Core Chiplet



Nap
(per core)

Core + L2
L3 cache

Stop clocks to only processor core execution engines.
Leave all caches running.
Saves ~ 10% power with ~ 5us Latency

Sleep
(per core)

Core + L2
L3 cache

Power OFF the core plus private L2 cache.
Requires restore/re-init to wakeup.
Leave shared L3 cache running.
Saves ~ 80% power with ~3ms Latency

Winkle
(per chiplet)

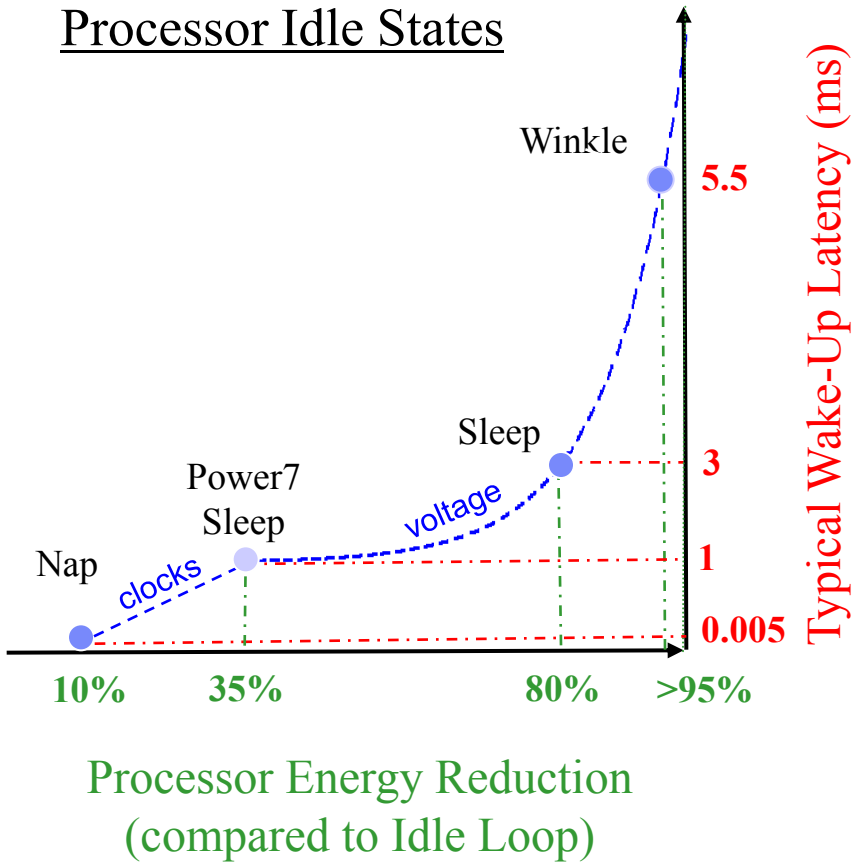
Core + L2
L3 cache

Power OFF the entire chiplet.
Requires restore/re-init to wakeup.
Takes offline 1/8 of the shared L3 cache.
Saves > 95% power with < 6ms Latency

Save Energy When Idle

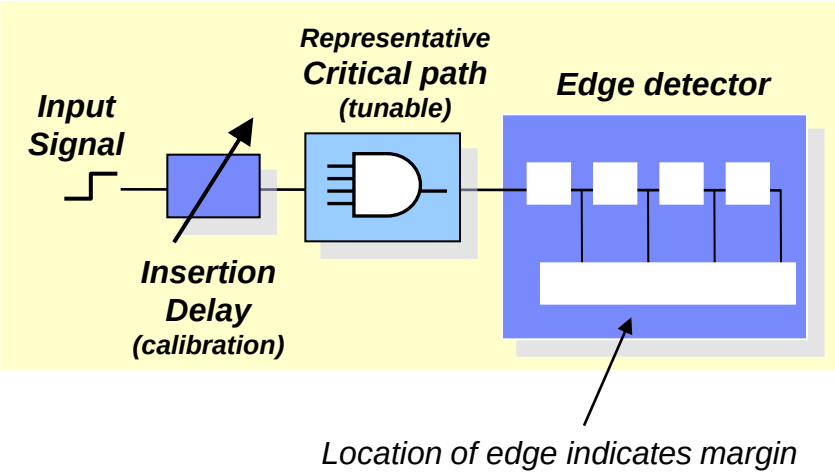
Three idle states were implemented to optimize power vs. latency

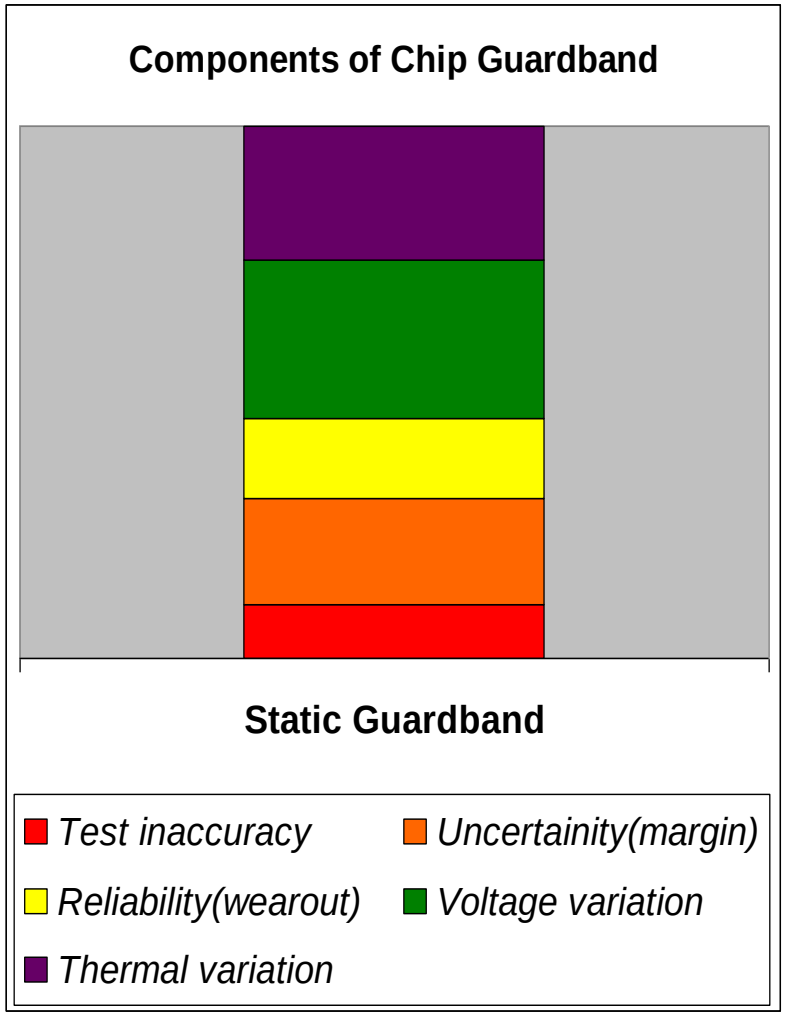
- **Nap** (Continued POWER7 support)
 - Optimized for wake-up time
 - Turn off clocks to execution units only
 - Caches remain coherent
- **Sleep** (Improved from POWER7)
 - More savings at increased latency
 - Purge and power off core plus L2 caches
 - Leave shared L3 cache running
- **Winkle** (New for POWER7+)
 - Maximum savings at higher latency
 - Purge and power off entire chiplet
 - Takes eighth of chip L3 cache offline



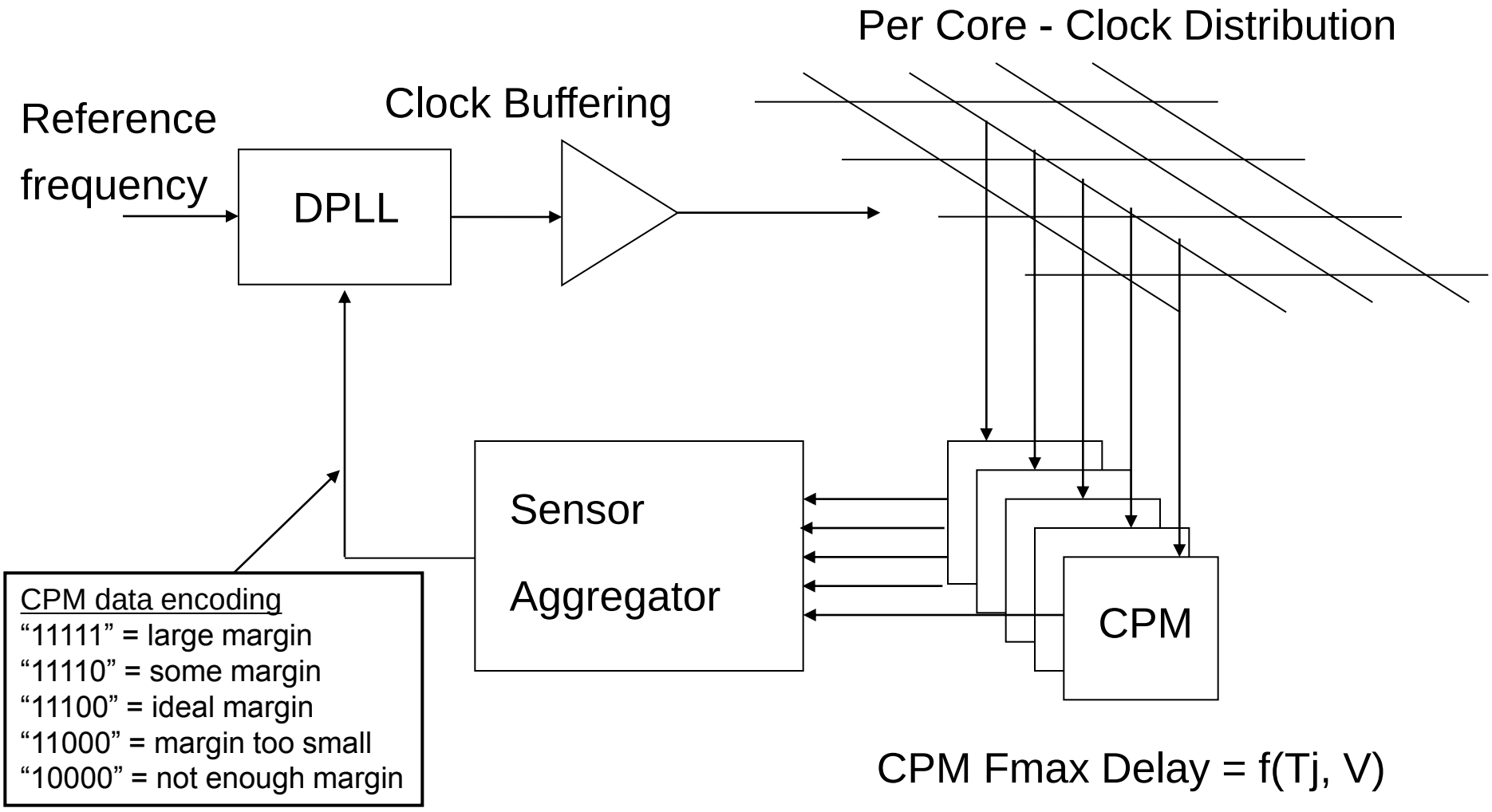
Real Time Chip Guardband

- **Conventional guardband**
 - Static, conservative voltage margins for potential worst-case conditions
 - Causes unnecessary loss of energy efficiency during typical server usage
- **Critical Path Monitor (CPM)**
 - Real Time detection of available circuit timing margin

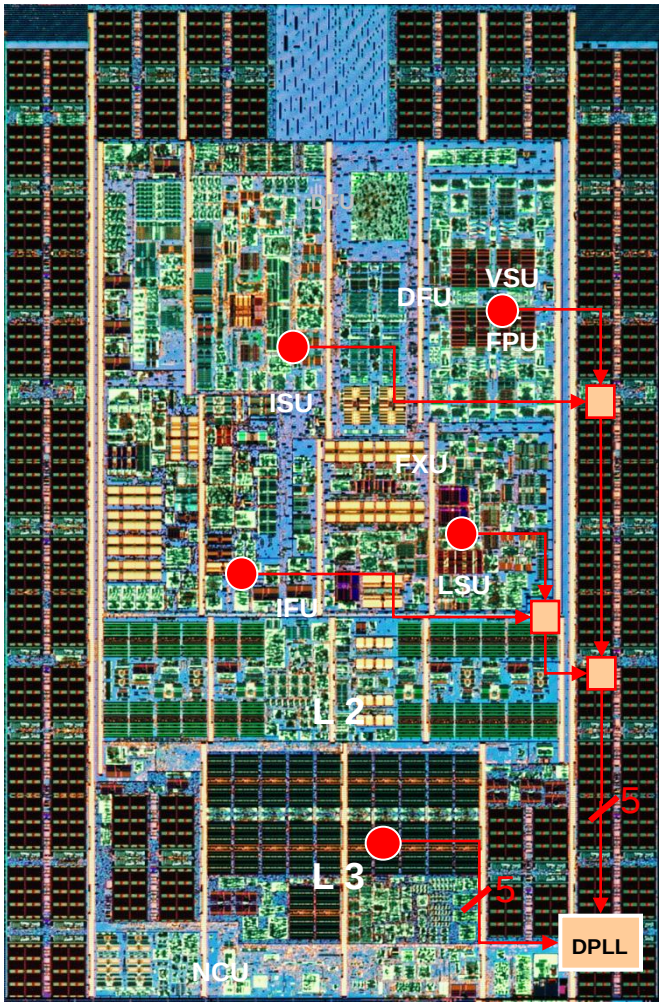




Real Time Guardband – DPLL/CPM feedback loop



POWER7+ Core CPM Infrastructure



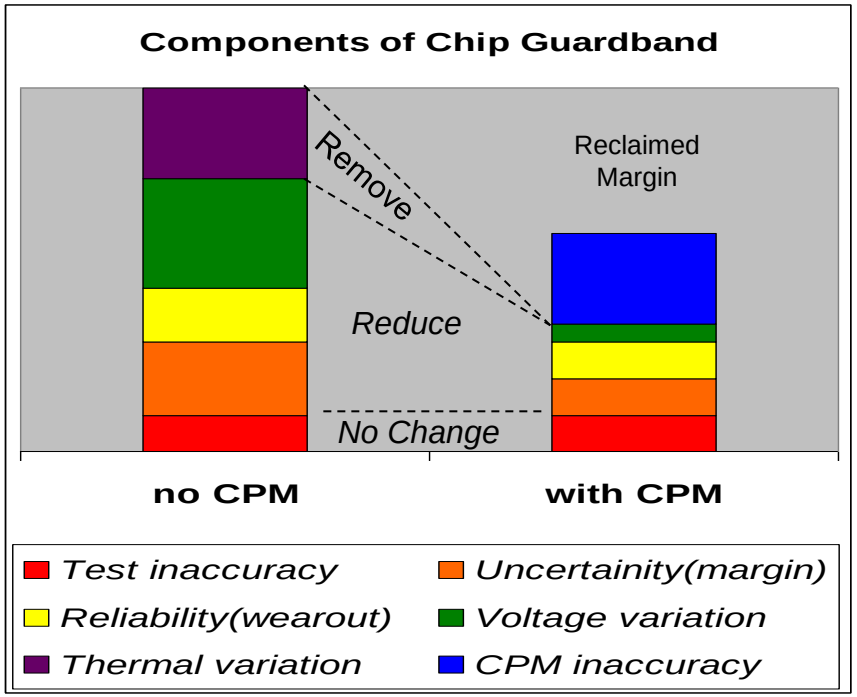
● = CPM (Critical Path Monitor)

▭ = AND Buffer

CPMs are strategically placed in known hot spots typically near micro-architecture critical paths.

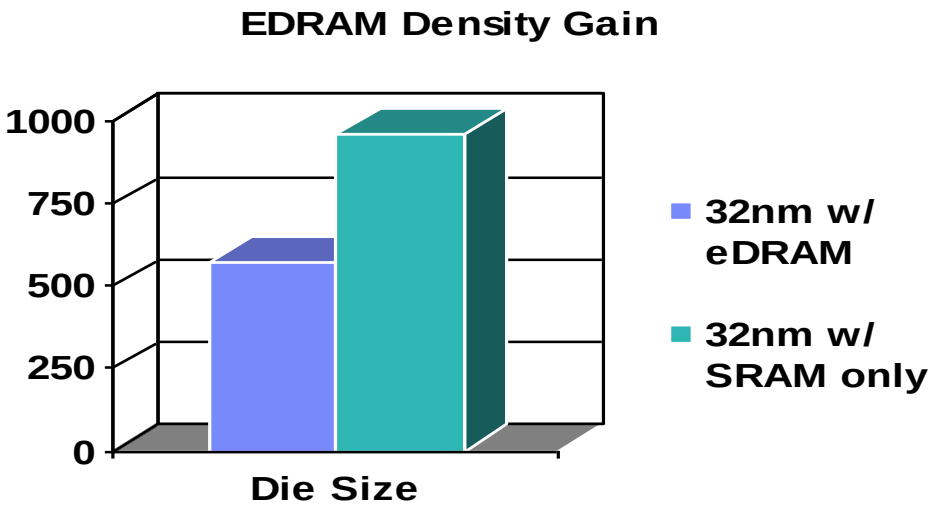
The real time feedback from CPMs can reduce how much margin is needed for various guardband components.

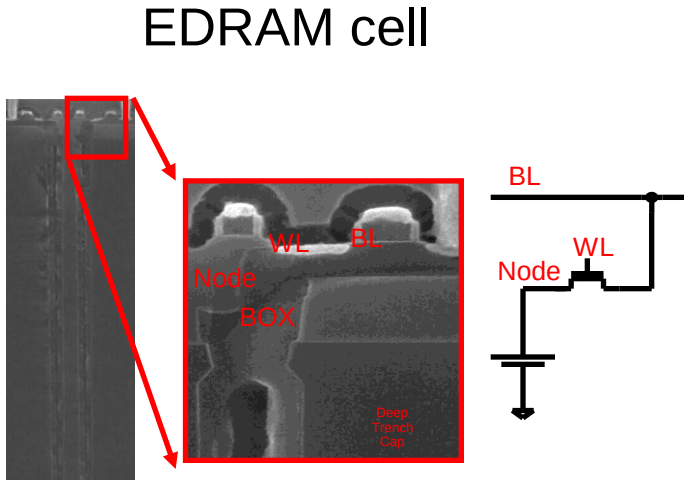
Real-time guardbanding will allow for greater energy efficiency.



Advanced 32nm Technology

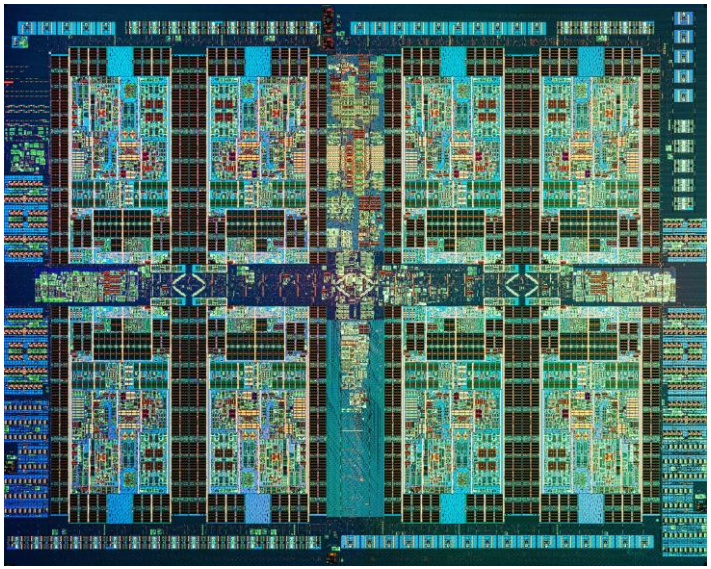
- 32nm High-K Metal Gate (HKMG) SOI based logic technology
 - 3 logic transistor threshold voltages (V_t) optimizes power/performance
 - 13-layer BEOL metal stack minimizes cross die latency
 - 1x, 2x, 4x, 8x, & ultra-thick metal layers
 - eDRAM provides 3-4x density advantage over SRAM
- Advanced features make the node effectively perform as one with sub-32nm features

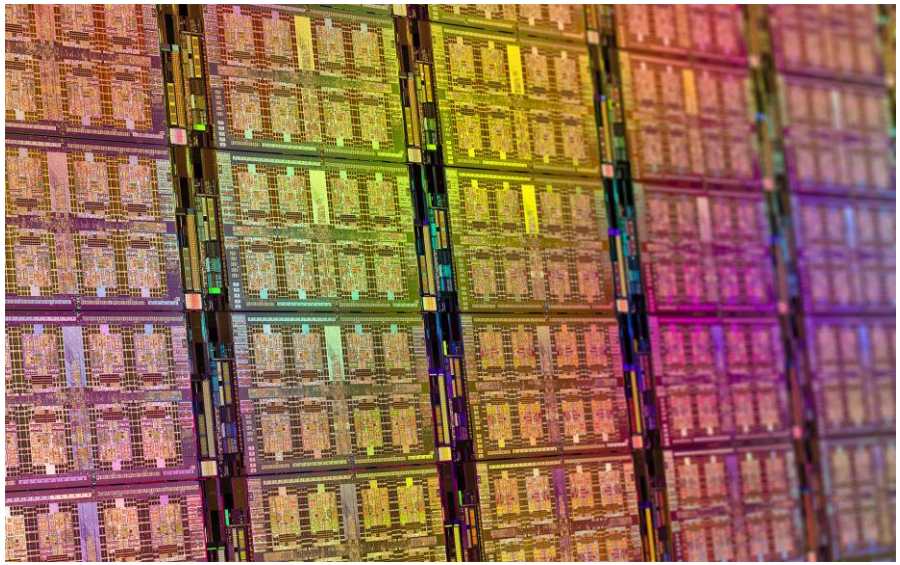




POWER7+: The next major step in IBM's roadmap

- Brings significant improvement to both scale up & scale out systems.
- The new accelerators optimize specific functions while offloading CPU.
- Advanced energy management greatly improves data center efficiency.





Acknowledgements:

Many Thanks to the entire POWER7+ design, manufacturing and product teams.

Questions: Send email to: sctaylor@us.ibm.com

The Intel® Xeon® Processor E5 Family

Architecture, Power Efficiency, and Performance



Jeff Gilbert, Sr PE
Mark Rowland, PE
August 2012

Agenda

1. Architecting Performance
2. Energy Efficiency from the Load Line to the Data Center
3. Measured Performance

Legal Disclaimer - Notice

- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

 UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

 Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined". Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.

 The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

 Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order.

 Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or go to: http://www.intel.com/design/literature.htm
- Intel processor numbers are not a measure of performance. Processor numbers differentiate features within each processor family, not across different processor families: Go to: http://www.intel.com/products/processor_number
- Intel® AES-NI requires a computer system with an AES-NI enabled processor, as well as non-Intel software to execute the instructions in the correct sequence. AES-NI is available on select Intel® processors. For availability, consult your reseller or system manufacturer. For more information, see http://software.intel.com/en-us/articles/intel-advanced-encryption-standard-instructions-aes-ni/
- No computer system can provide absolute security under all conditions. Intel® Trusted Execution Technology (Intel® TXT) requires a computer with Intel® Virtualization Technology, an Intel TXT-enabled processor, chipset, BIOS, Authenticated Code Modules and an Intel TXT-compatible measured launched environment (MLE). Intel TXT also requires the system to contain a TPM v1.s. For more information, visit http://www.intel.com/technology/security
- Intel® Virtualization Technology requires a computer system with an enabled Intel® processor, BIOS, and virtual machine monitor (VMM). Functionality, performance or other benefits will vary depending on hardware and software configurations. Software applications may not be compatible with all operating systems. Consult your PC manufacturer. For more information, visit http://www.intel.com/go/virtualization
- Requires a system with Intel® Turbo Boost Technology. Intel Turbo Boost Technology and Intel Turbo Boost Technology 2.0 are only available on select Intel® processors. Consult your PC manufacturer. Performance varies depending on hardware, software, and system configuration. For more information, visit http://www.intel.com/go/turbo
- Intel product is manufactured on a lead-free process. Lead is below 1000 PPM per EU RoHS directive (2002/95/EC, Annex A). No exemptions required
- Halogen-free: Applies only to halogenated flame retardants and PVC in components. Halogens are below 900ppm bromine and 900ppm chlorine.
- Copyright © 2012 Intel Corporation. All rights reserved. Intel, Intel Xeon, the Intel Xeon logo and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries. .
*Other names and brands may be claimed as the property of others.

Legal Disclaimers - Performance

- Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
- Intel does not control or audit the design or implementation of third party benchmarks or Web sites referenced in this document. Intel encourages all of its customers to visit the referenced Web sites or others where similar performance benchmarks are reported and confirm whether the referenced benchmarks are accurate and reflect performance of systems available for purchase.
- Relative performance is calculated by assigning a baseline value of 1.0 to one benchmark result, and then dividing the actual benchmark result for the baseline platform into each of the specific benchmark results of each of the other platforms, and assigning them a relative performance number that correlates with the performance improvements reported.
- SPEC, SPECint, SPECfp, SPECrate, SPECpower_ssj, SPECjAppServer, SPECjEnterprise, SPECjbb, SPECompM, SPECompL, and SPEC MPI are trademarks of the Standard Performance Evaluation Corporation. See http://www.spec.org for more information.
- TPC Benchmark is a trademark of the Transaction Processing Council. See http://www.tpc.org for more information.
- SAP and SAP NetWeaver are the registered trademarks of SAP AG in Germany and in several other countries. See http://www.sap.com/benchmark for more information.

Optimization Notice

Optimization Notice

Intel® compilers, associated libraries and associated development tools may include or utilize options that optimize for instruction sets that are available in both Intel® and non-Intel microprocessors (for example SIMD instruction sets), but do not optimize equally for non-Intel microprocessors. In addition, certain compiler options for Intel compilers, including some that are not specific to Intel micro-architecture, are reserved for Intel microprocessors. For a detailed description of Intel compiler options, including the instruction sets and specific microprocessors they implicate, please refer to the "Intel® Compiler User and Reference Guides" under "Compiler Options." Many library routines that are part of Intel® compiler products are more highly optimized for Intel microprocessors than for other microprocessors. While the compilers and libraries in Intel® compiler products offer optimizations for both Intel and Intel-compatible microprocessors, depending on the options you select, your code and other factors, you likely will get extra performance on Intel microprocessors.

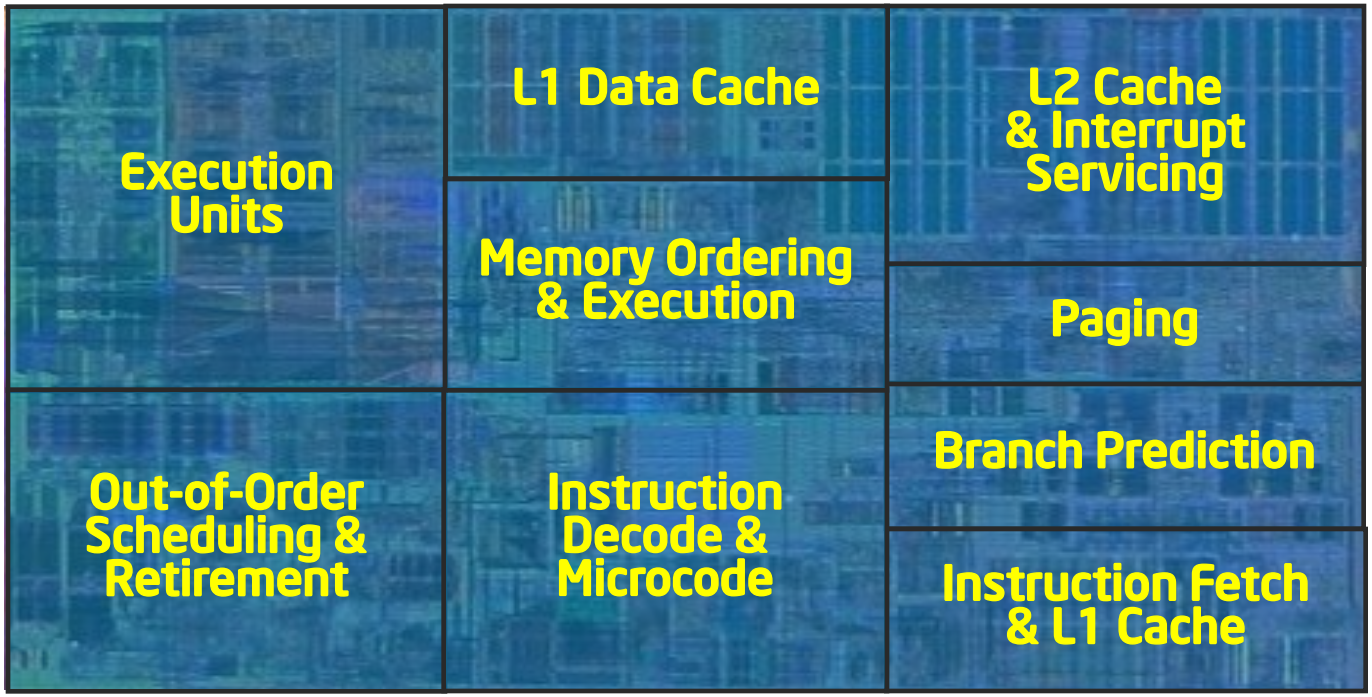
Intel® compilers, associated libraries and associated development tools may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include Intel® Streaming SIMD Extensions 2 (Intel® SSE2), Intel® Streaming SIMD Extensions 3 (Intel® SSE3), and Supplemental Streaming SIMD Extensions 3 (Intel® SSSE3) instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors.

While Intel believes our compilers and libraries are excellent choices to assist in obtaining the best performance on Intel® and non-Intel microprocessors, Intel recommends that you evaluate other compilers and libraries to determine which best meet your requirements. We hope to win your business by striving to offer the best performance of any compiler or library; please let us know if you find we do not.

Notice revision #20101101

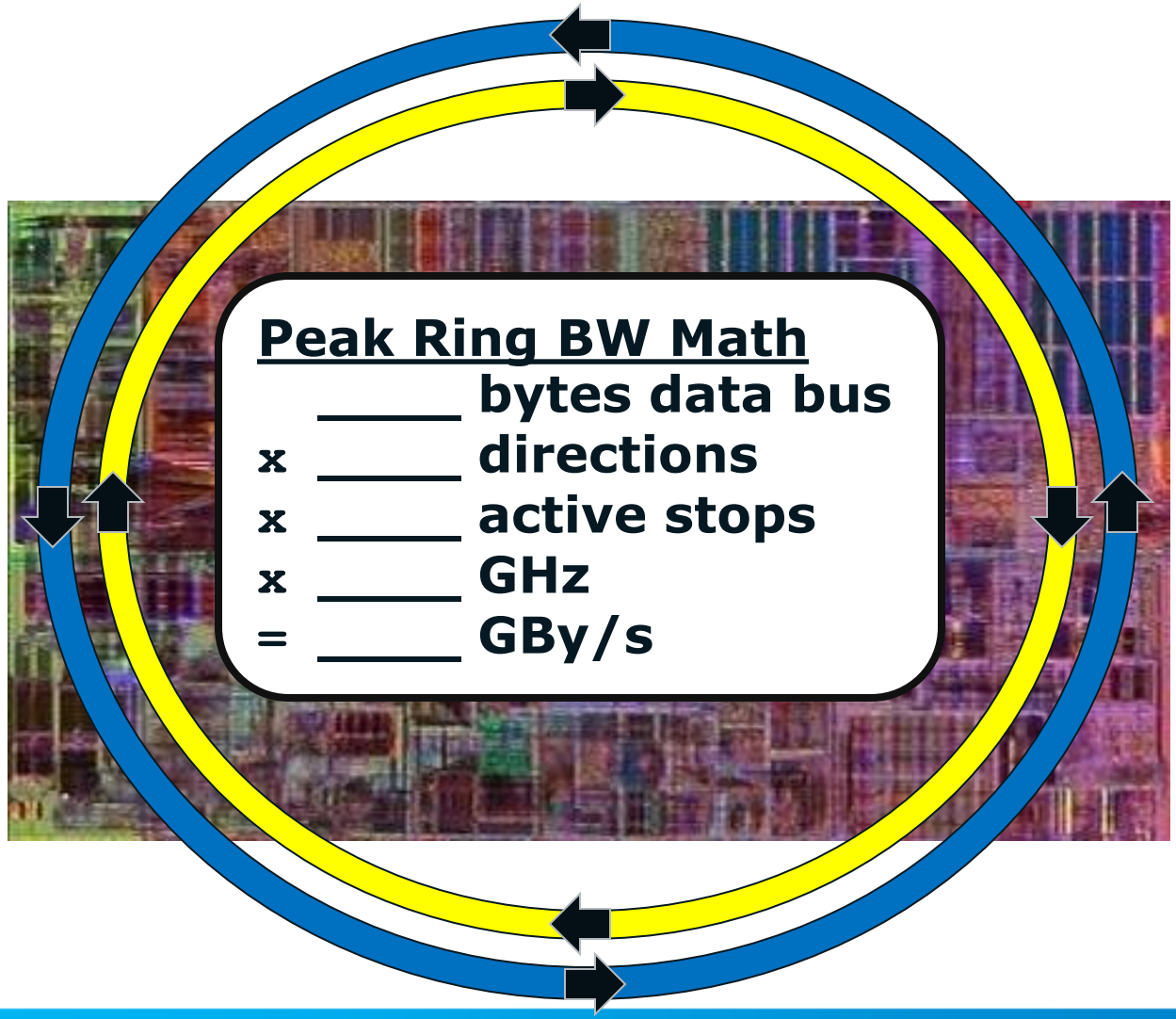
Foundations of SNB-EP Performance

Start with the Sandy Bridge Core



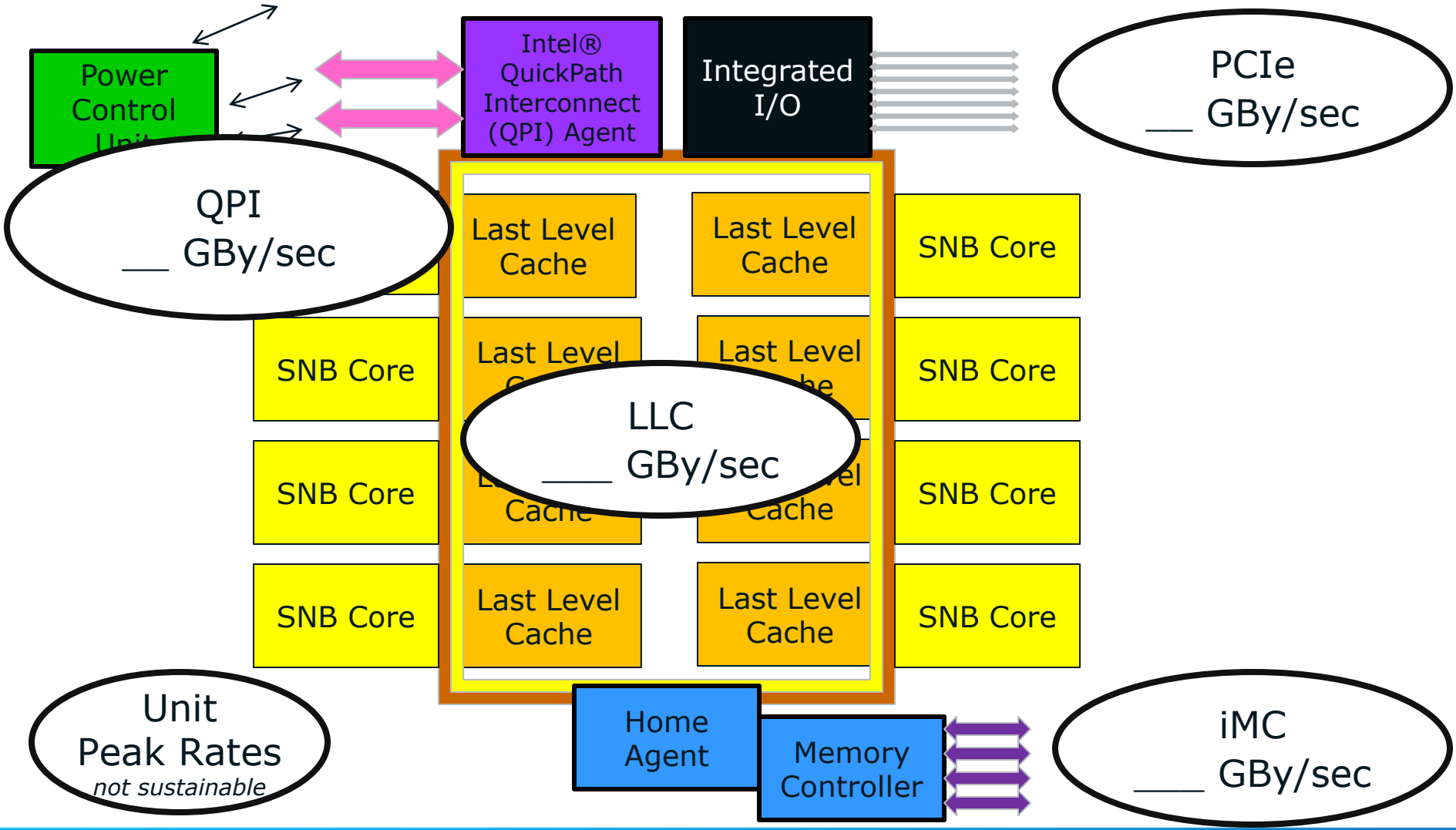
Foundations of SNB-EP Performance

Put Eight Cores on a High BW Interconnect: The Ring

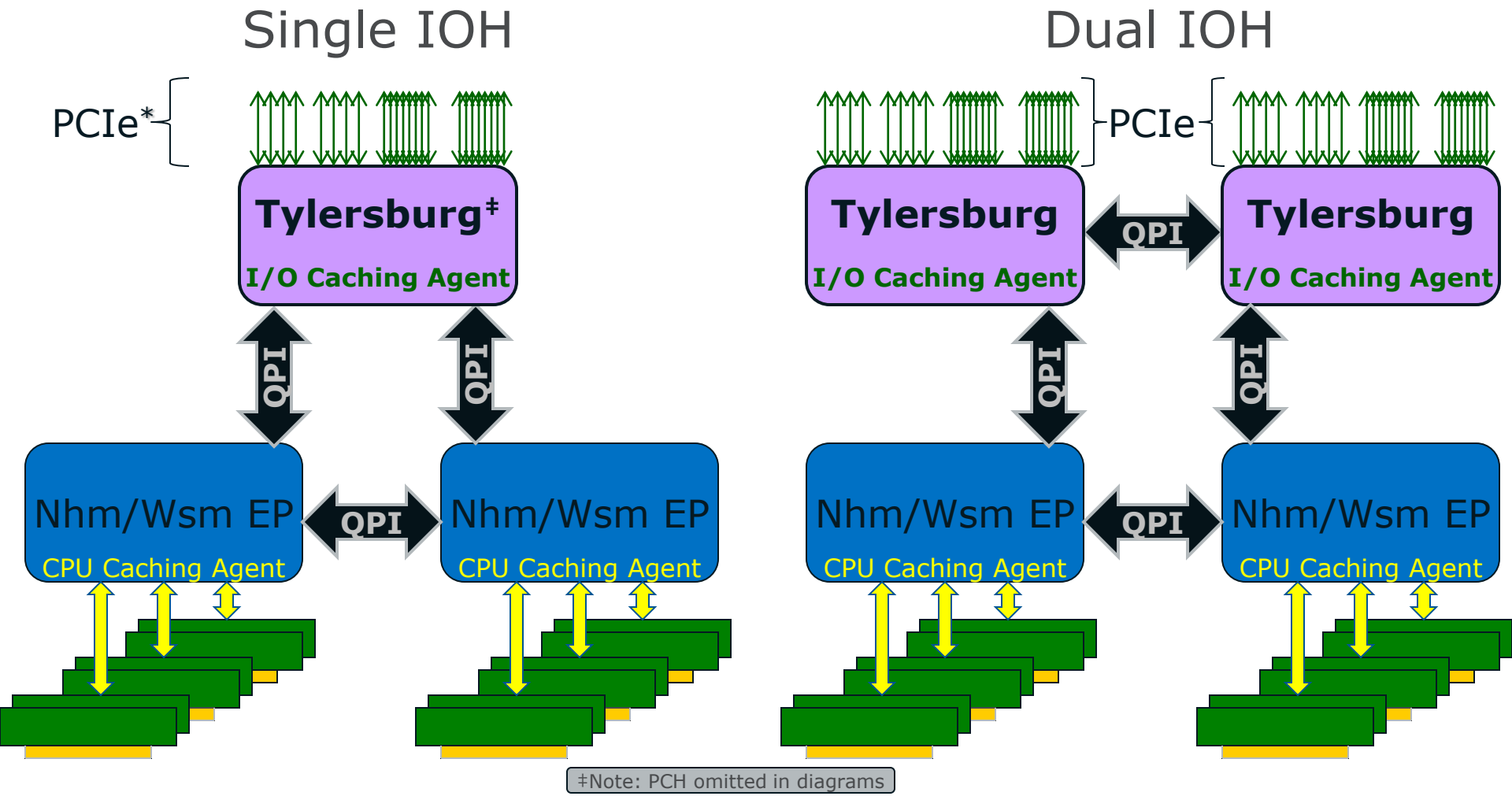


Foundations of SNB-EP Performance

Add an LLC, System Agents, and Power Management



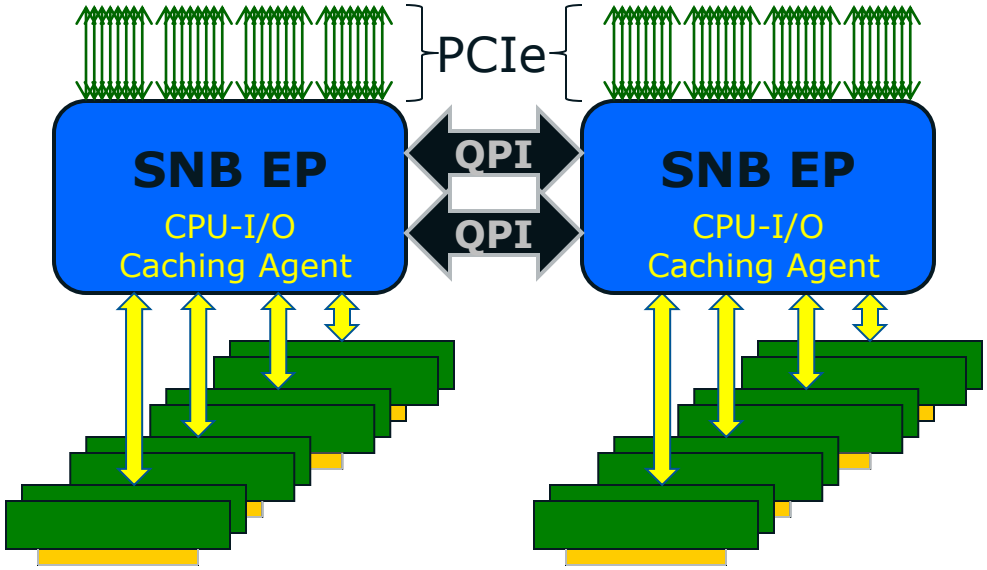
Thurley Platform Review



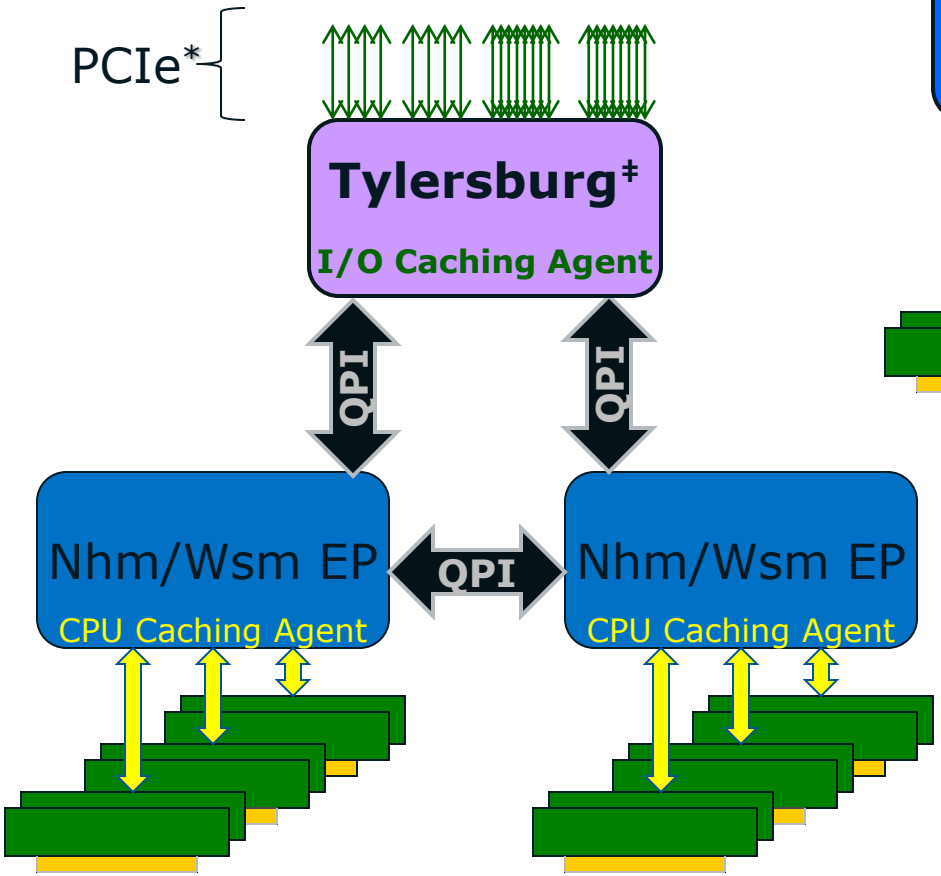
Foundations of SNB-EP Performance

Higher Performance Platform

Romley



Thurley



Topology Performance Changes

- 40 Lanes of 8 GT/s Integrated PCIe
- Dual Inter-processor QPI links
- Four higher speed memory channels

‡Note: PCH omitted in diagrams

Foundations of SNB-EP Performance

Focus on I/O Performance

- **PCIe G3: 8 GT/s vs. 4 GT/s**
 - DMI2 (4 GT/s) vs. DMI1 (2 GT/s) (not shown in diagram)
- **I/O capacity scales with sockets (memory BW)**
- **Inherent benefit from Integration:**
 QPI link to I/O controller replaced with direct ring interconnect reducing latency and increasing BW
- **CPUs and PCIe are a unified Caching Agent**
 - Less resource partitioning
 - More scalable, higher performance
 - Reduces the latency of cacheable traffic
 - PCIe acts under the auspices of and uses the LLC (more later)

Foundations of SNB-EP Performance

Focus on I/O Performance (cont'd)

- **I/O-related Optimizations**
 - Double width data buses in the I/O unit
 - ReadCurrent semantics rather the Code Read
 - Potentially reduces memory write traffic – maybe a lot
 - Inbound writes
 - Cache line pre-allocated but ownership can be preempted
 - Prefetch of data (for write merging)
- **40 lanes vs. 36 lanes**
- **Physical address range (46b vs. 41b)**

Foundations of SNB-EP Performance

Focus on I/O Performance (cont'd)

- **Intel® Direct Data I/O Technology (Intel® DDIO): IIO allocates and transfers directly into LLC**
 - IIO cache allocating is generally limited to 2 (of 20) ways
 - Can use a line that's already been allocated by, say, a core
 - Circular buffers of reasonable size (a few to ten MBy) can reside in the LLC *and, in practice, almost never be written*.
 - Making use of this can effectively double the achievable I/O bandwidth of a core and of a socket.
 - Permits practically linear scaling as multiple high bandwidth I/O devices are added (e.g., 10 GbE adapters) with achieving nearly zero read and write bandwidth to memory
 - Saves power, too

Mid-Game Summary

- **Improved performance by improving the parts**
 - Sandy Bridge core
 - On-die interconnect (“Ring”)
 - More and faster memory channels with improved scheduling
 - Faster inter-socket communication (Intel® QPI)
 - Integrating and accelerating I/O

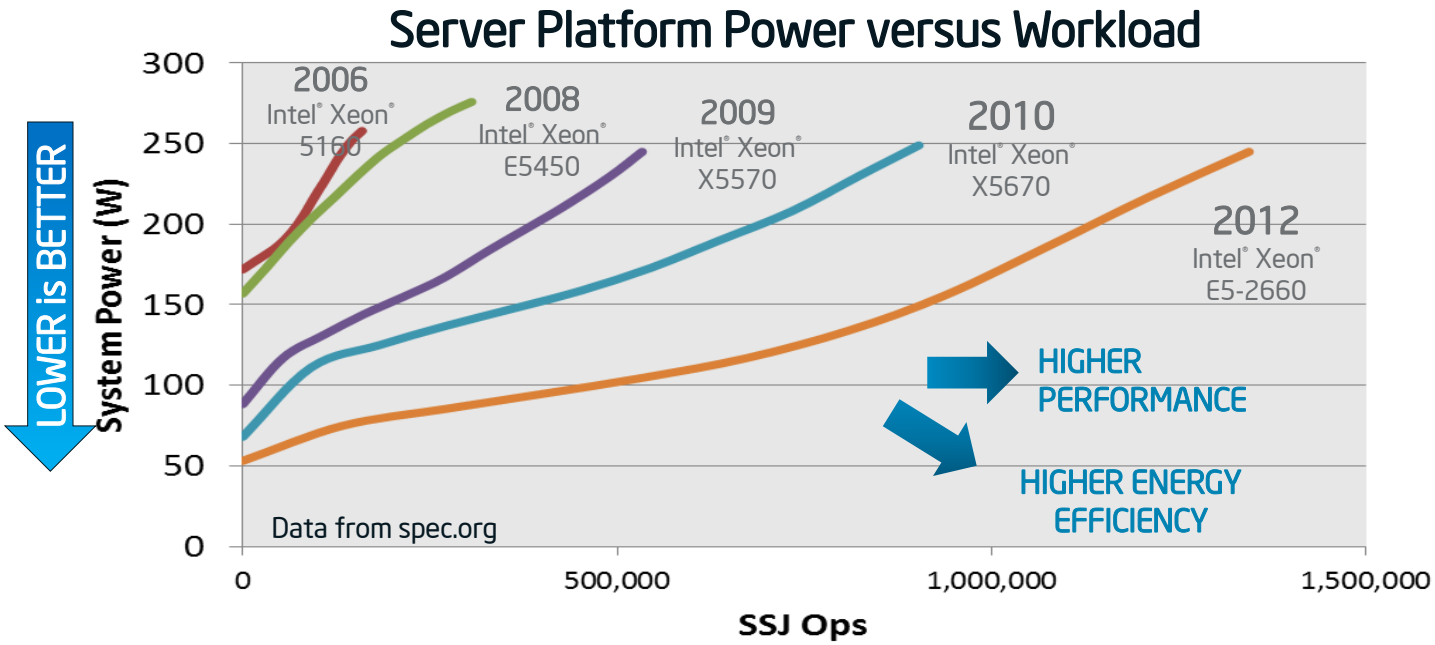
- **Coming Up in the Next Half:**

 Performance with Power Efficiency

Energy Efficient Load Line



Energy Efficient Load Line



Performance:
45% CAGR

Energy Efficiency:
60% CAGR

- **Platform efficiency at low Power**
 - CPU and DRAM VR Phase shedding
- **Scalable Uncore Power**
 - Uncore voltage frequency scaling
- **Scalable Memory Power**
 - Multi-rank slow CKE
- **Processor Power**
 - Energy Perf BIAS, Dynamic Switching
- **I/O Power management**
 - QPI L0p/L1, PCIE ASPM L1

Significant Improvement to Proportional Energy

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. Configurations: Dual Socket Server. For full configuration information, please see backup. For more information go to http://www.intel.com/performance

Dynamic Performance Load Line

PCU dynamically adjusts to OS Power Management Policy

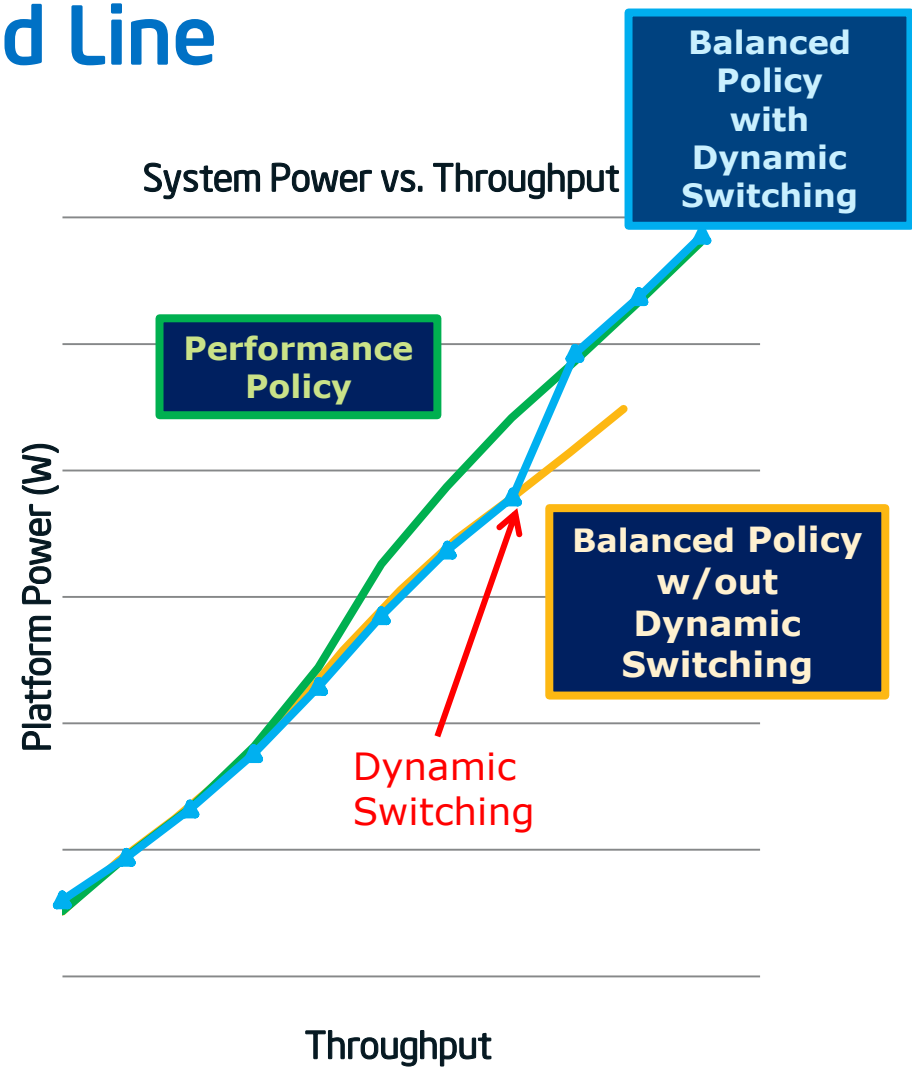
- OS communicates Policy through EPB (Energy Perf BIAS)
- PCU monitors and adjusts autonomous on die power saving engines

PCU automatically adjusts for Performance at high utilization

- Leverages EPB to switch into performance mode when necessary

Optimized across a range of workloads

- Single-threaded workloads
- Multi-threaded workloads

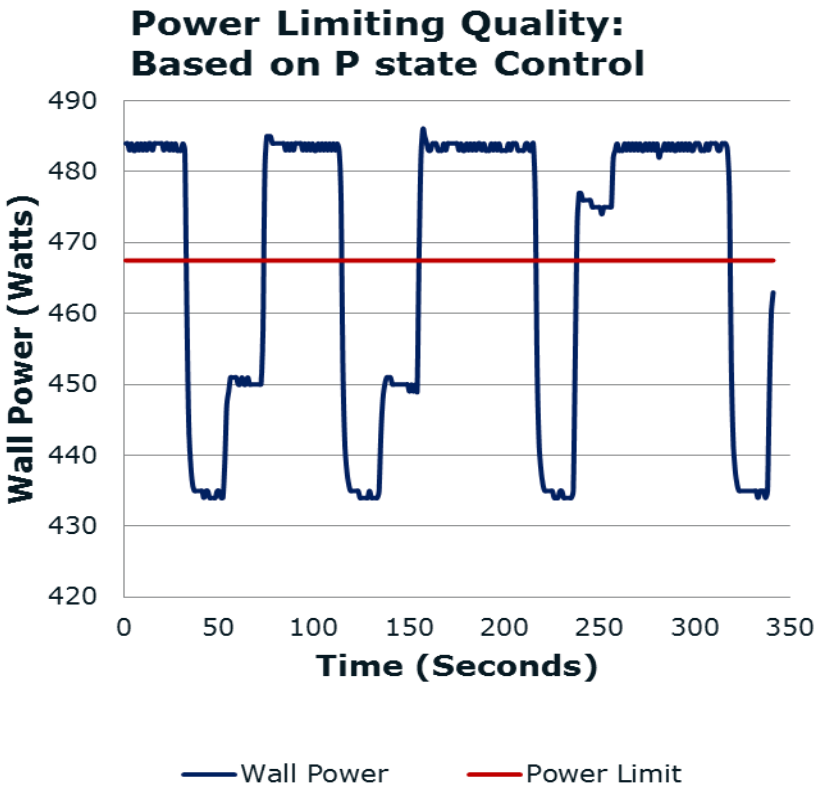


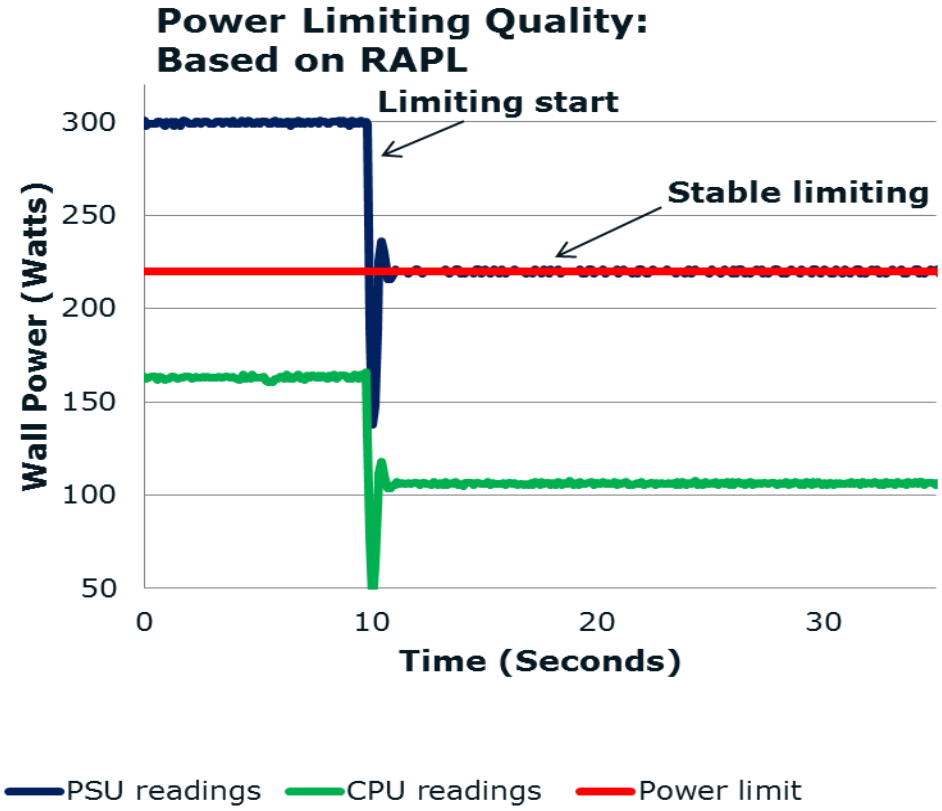
PCU works synergistically with OS Power Policy

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark* and MobileMark*, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. Configurations: Dual Socket Server. For full configuration information, please see backup. For more information go to http://www.intel.com/performance


Energy Efficiency in the Data Center

Running Average Power Limiting (RAPL)

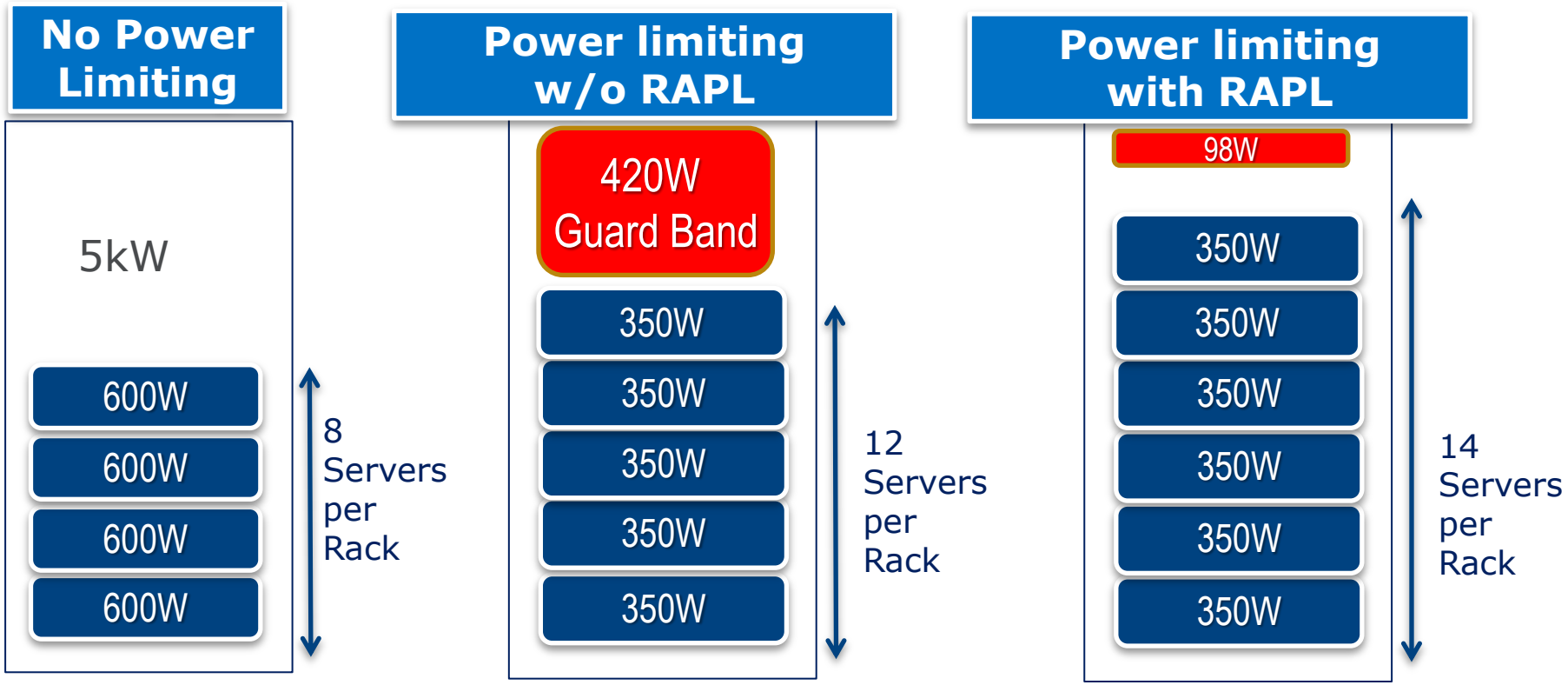




RAPL gives accurate and stable power limiting than P state control

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

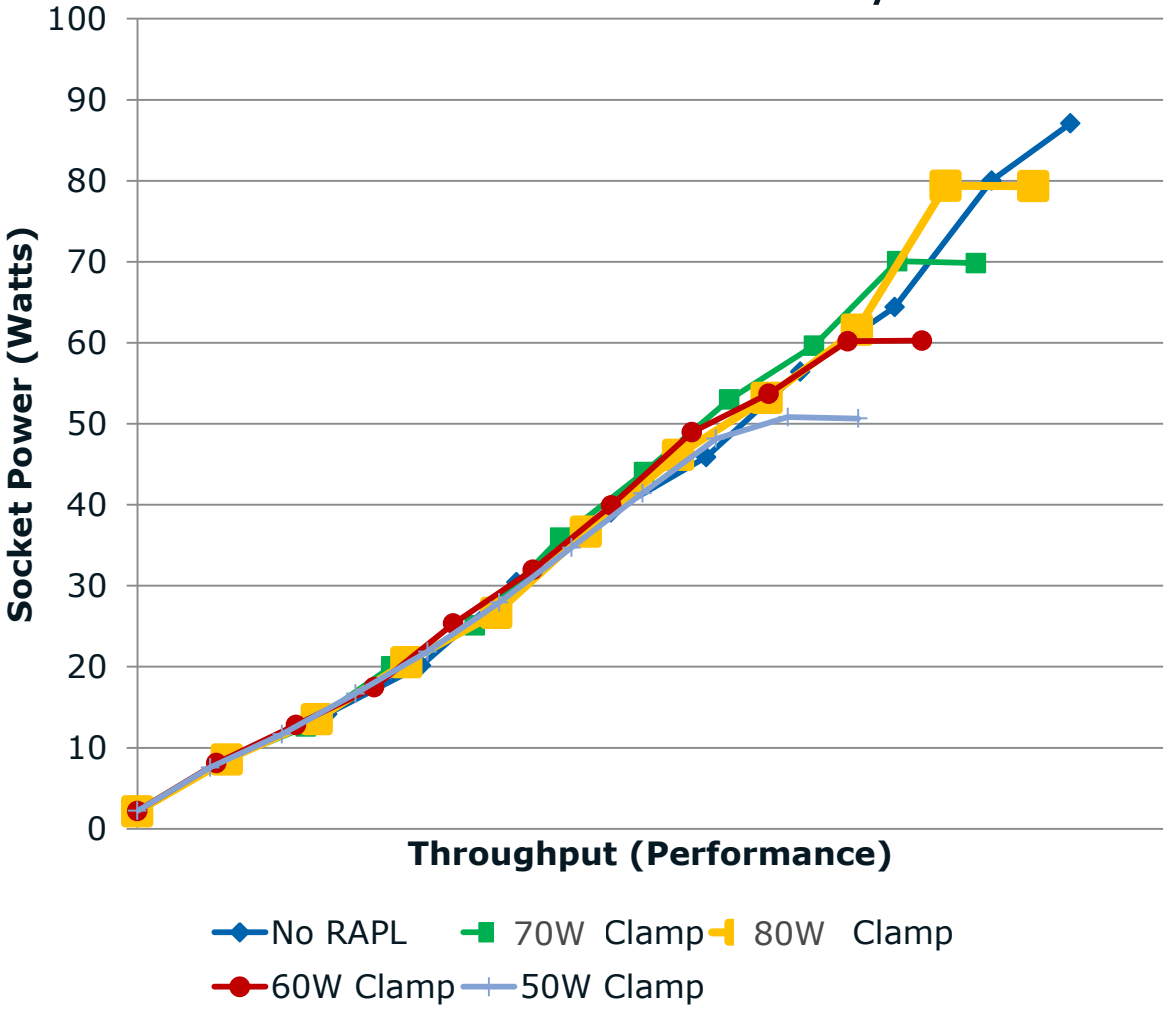
Improved Efficiency with RAPL



Max System Power: 600W
Typical System Power: 350W
Rack Power: 5kW

Improved Power Limiting Accuracy Allows for Smaller Guard bands and Increased Rack Density.

Socket RAPL & the Power/Performance Load Line



- **Power Clamps effectively limit socket power**
- **Max possible performance achieved for given power limit**
- **Efficient Power Scaling across the load line**

Internal Estimates
Q2 2011
SNB EP, C0 stepping

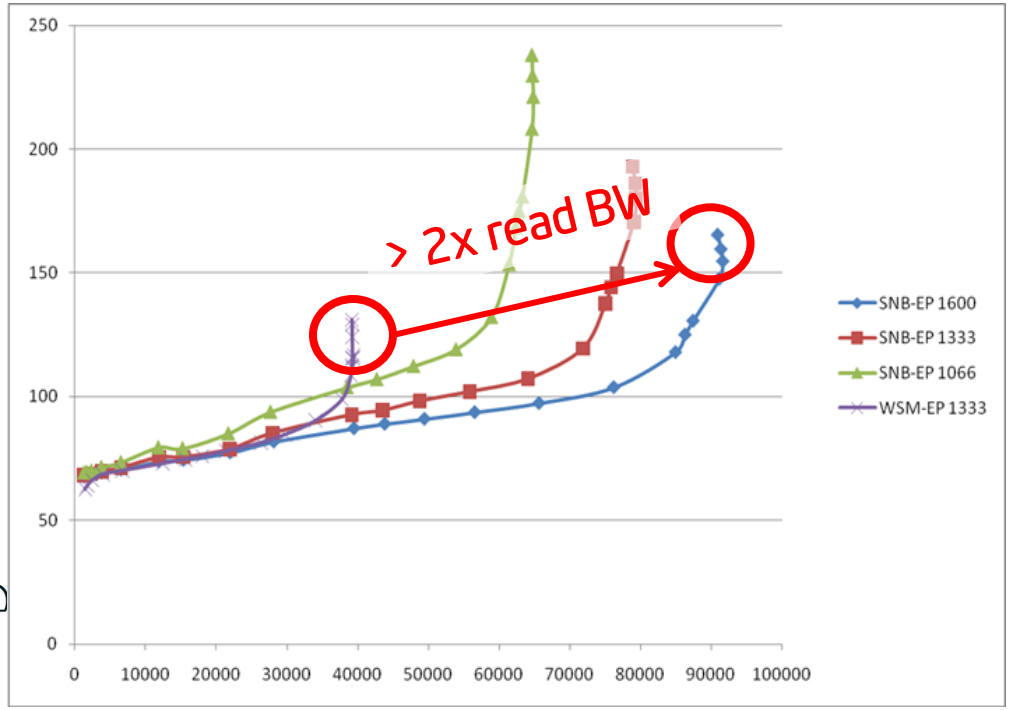
Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Maximum Performance



Memory Latency Optimizations

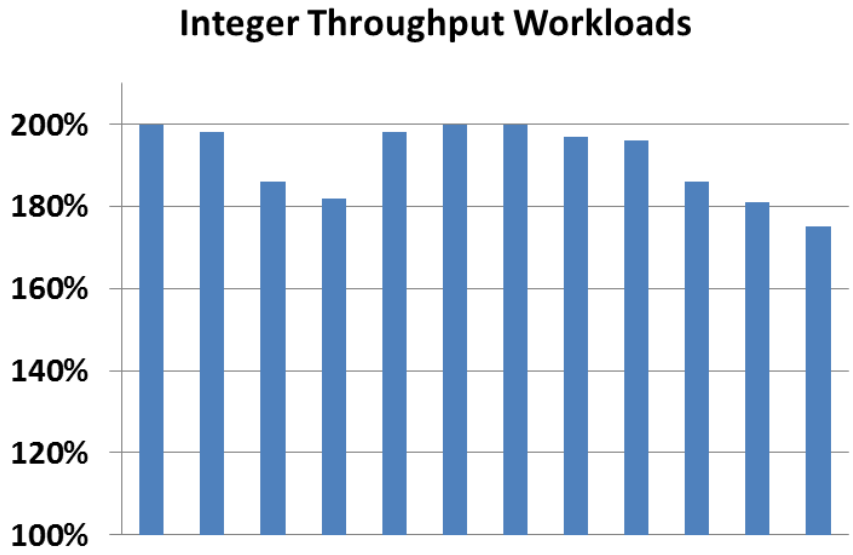
- Early Snoop
 Dynamic Direct 2 Core
- Uncore Frequency Change
- Dynamic Memory CKE Disable
- New LLC Prefetcher
- **Distributed L3**
 - Theoretical Peak: ~844GB/s (1s @ 3.3GHz w/ 8 cores)
 - Core->L3 Read Throughput: >250GB/s (1s @ 3.3GHz w/ 8 cores)
- **Dual Load Ports on L1 D-Cache**
- **SandyBridge Turbo 2.0**

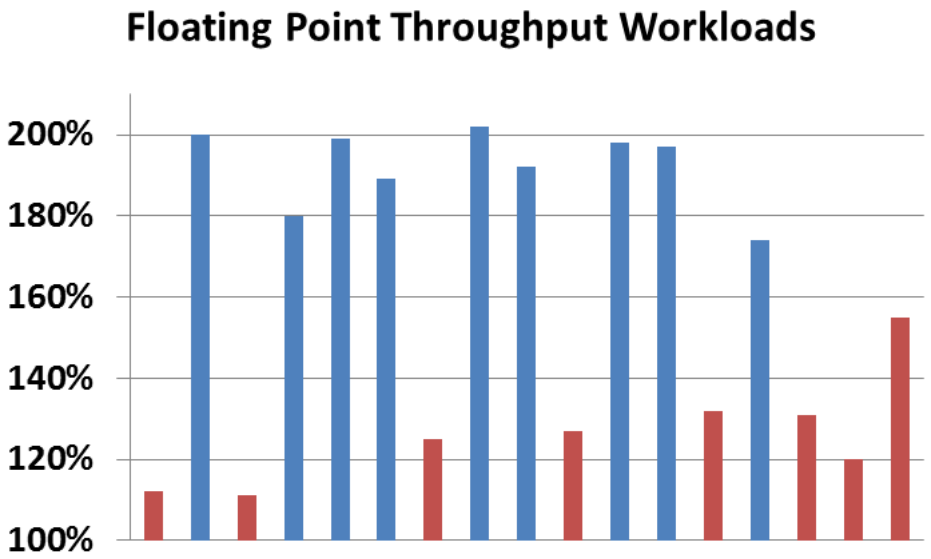


>2x max bandwidth from Xeon 5600 on read BW
- 3->4 channels (+33%)
- 1333->1600 (+20%)
- Improved Efficiency (+~40%)

Benchmark Notes:
- Intel internal tool for BW and Latency

Comparison of 4 core to 8 core Scaling @ 3.3GHz





- Core sensitive apps in both INT and FP show excellent performance scaling
- Memory sensitive apps show less scaling (as expected shown in red)

Internal Testing – Estimate
4c: SNB E5-2643 w/out Turbo (1 DPC, DDR 1600)
8c: SNB E5-2690 w/ Turbo (2 DPC, DDR 1600)
ICC 12.1 / RHEL 6.1 / 2.6.32.131

Apps highlighted in Red are Memory Bandwidth sensitive

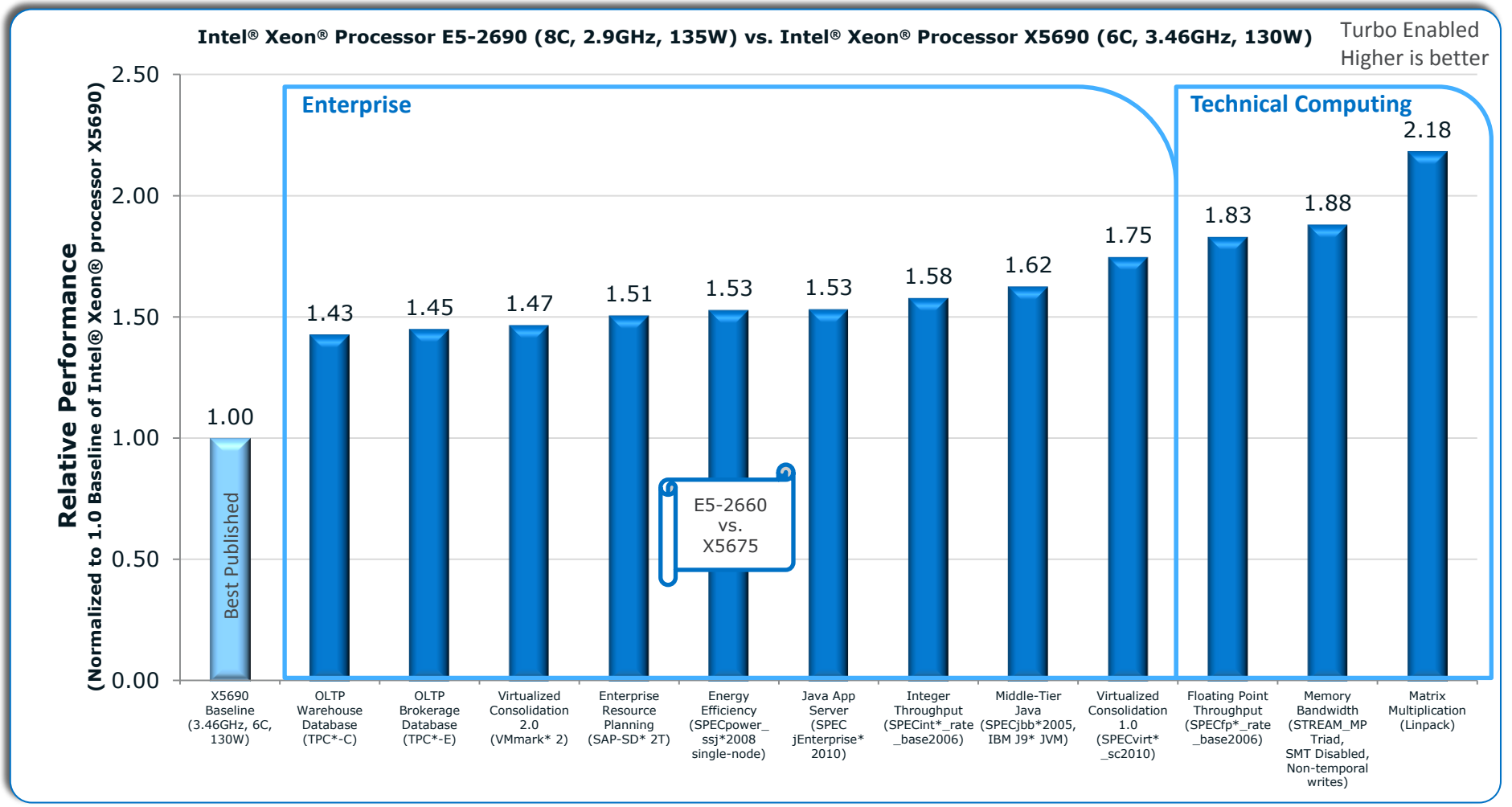
Intel® Xeon® E5 uncore provides significant core Scaling

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Configuration Details for Foil #25

For the SPEC benchmarks, please see http://www.spec.org for more information
Configuration Details: As of 31 May 2012
SAP* SD 2-tier
2x Intel Xeon processor X5690 (12M cache, 3.46GHz, 6.40GT/s Intel QPI) score 5220 SD users. Certification #2011005. Source: http://download.sap.com/download.epd?context=40E2D9D5E00EEF7C4B299992CE278ECED5166ED278FF20DF78759DC5B1E5FE79;
2x Intel Xeon processor E5-2690 (20M cache, 2.9GHz, 8.0GT/s Intel QPI) score 7865 SD users. Source: http://download.sap.com/download.epd?context=40E2D9D5E00EEF7C5DDB3927818D671E00ECF023B5CE29EE68B565E9F19F1254
SPECvirt_sc*2011
2x Intel® Xeon® processor X5690 (6C,12M, 3.06GHz) score 1367 @ 84 VMs. Source: http://www.spec.org/virt_sc2010/results/res2011q1/virt_sc2010-20110209-00022-perf.html;
2x Intel® Xeon® processor E5-2690 (8C, 2.9GHz, C0) score 2,388 @ 150 VMs. Source: http://www.spec.org/virt_sc2010/results/res2012q2/virt_sc2010-20120403-00045-perf.html
SPECpower_ssj*2008
metrics for SPECpower are efficiency based and expressed as ssj_ops/watt.
2x Intel Xeon processor X5675 (12M cache, 3.45GHz, 6.40GT/s Intel QPI) score 3,329. Source: http://www.spec.org/power_ssj2008/results/res2011q4/power_ssj2008-20110713-00386.html;
2x Intel Xeon processor E5-2660 (20M cache, 2.2GHz, 8.0GT/s Intel QPI, C1) score 5,088. Source: http://www.spec.org/power_ssj2008/results/res2012q2/power_ssj2008-20120427-00454.html
TPC-E*
2x Intel Xeon processor X5690 (12M Cache, 3.46GHz, 2P/12C/24T) referenced as published at 1,284.14 tpsE, $250 USD/tpsE, available 5/4/11. Source: http://www.tpc.org/tpce/results/tpce_result_detail.asp?id=111050403;
Intel: 2x Intel Xeon processor E5-2690 (20M cache, 2.9GHz, 2P/16C/32T) referenced as published at 1,863.23 tpsE, $207.85 USD/tpsE, available 3/6/12. Source: http://www.tpc.org/tpce/results/tpce_result_detail.asp?id=112030601
VMmark* 2
2x Intel Xeon processor X5690 (12M cache, 3.46GHz, 6.40GT/s Intel QPI) score 7.59 @ 7 Tiles. Source: http://www.vmware.com/a/assets/vmmark/pdf/2011-10-18-Fujitsu-RX300S6.pdf;
2x Intel Xeon processor E5-2690 (20M cache, 2.9GHz, 8.0GT/s Intel QPI, C1) score 11.13 @ 10 Tiles. Source: http://www.vmware.com/a/assets/vmmark/pdf/2012-05-15-HP-DL360pG8.pdf
TPC-C*
2x Intel Xeon processor X5690 (12M Cache, 3.46GHz, 2P/12C/24T) referenced as published at 1,053,100 tpmC, $0.57 USD/tpmC, available 6/20/11. Source: http://www.tpc.org/tpcc/results/tpcc_result_detail.asp?id=111120802;
2x Intel Xeon processor E5-2690 (20M cache, 2.9GHz, 8.0GT/s Intel QPI) referenced as published at 1,503,544 tpmC, $0.53 USD/tpmC, available 4/11/12. Source: http://www.tpc.org/tpcc/results/tpcc_result_detail.asp?id=112041101
SPECjbb*2005
2x Intel Xeon processor X5690 (12M cache, 3.46GHz, 6.40GT/s Intel QPI) score 975,257 bops, 487,629 bops/JVM. Source: http://www.spec.org/osg/jbb2005/results/res2011q1/jbb2005-20110215-00950.html;
2x Intel Xeon processor E5-2690 (2.9GHz, 8C) score 1,584,567 bops. Source: http://www.spec.org/osg/jbb2005/results/res2012q1/jbb2005-20120306-01056.html
SPECint*_rate_base2006
2x Intel Xeon processor X5690 (12M cache, 3.45GHz, 6.40GT/s Intel QPI) baseline score 425. Source: http://www.spec.org/cpu2006/results/res2012q2/cpu2006-20120322-20154.html
2x Intel Xeon processor E5-2690 (20M cache, 2.9GHz, 8.0GT/s Intel QPI) baseline score 671. Source: http://www.spec.org/cpu2006/results/res2012q1/cpu2006-20120307-19618.html
SPECjEnterprise*2010
2x Intel Xeon processor X5690 (12M cache, 3.46GHz, 6.40GT/s Intel QPI) score 5,427 EjOPS. Source: http://www.spec.org/jEnterprise2010/results/jEnterprise2010.html;
2x Intel Xeon processor E5-2690 (20M cache, 2.9GHz, 8.0GT/s Intel QPI) score 8,310.19 EjOPS. Source: http://www.spec.org/jEnterprise2010/results/jEnterprise2010.html
SPECfp*_rate_base2006
2x Intel Xeon processor X5690 (12M cache, 3.45GHz, 6.40GT/s Intel QPI) baseline score 271. Source: http://www.spec.org/cpu2006/results/res2012q1/cpu2006-20111219-19195.html
2x Intel Xeon processor E5-2690 (20M cache, 2.9GHz, 8.0GT/s Intel QPI) baseline score 496. Source: http://www.spec.org/cpu2006/results/res2012q1/cpu2006-20120307-19617.html
STREAM*_MP Triad (NTW)
2x Intel Xeon processor X5690 (12M cache, 3.45GHz, 6.40GT/s Intel QPI) TRIAD score 42GB/s. Source: Intel TR#1241
2x Intel Xeon processor E5-2690 (20M cache, 2.9GHz, 8.0GT/s Intel QPI, C1) score 79.5 GB/s. Source: Intel TR#1241
Linpack
2x Intel Xeon X5690 (12M cache, 3.45GHz, 6.40GT/s Intel QPI) score 159.4. Source: Intel TR#1236
2x Intel Xeon processor E5-2690 (20M cache, 2.9GHz, 8.0GT/s Intel QPI, C1) score 347.7. Source: Intel TR#1236
SPEC, SPECpower_ssj, SPECjEnterprise, SPECint, SPECjbb, SPECvirt_sc, and SPECfp are trademarks of SPEC

Intel® Xeon® Processor E5-2600 Product Family Generational Performance Summary



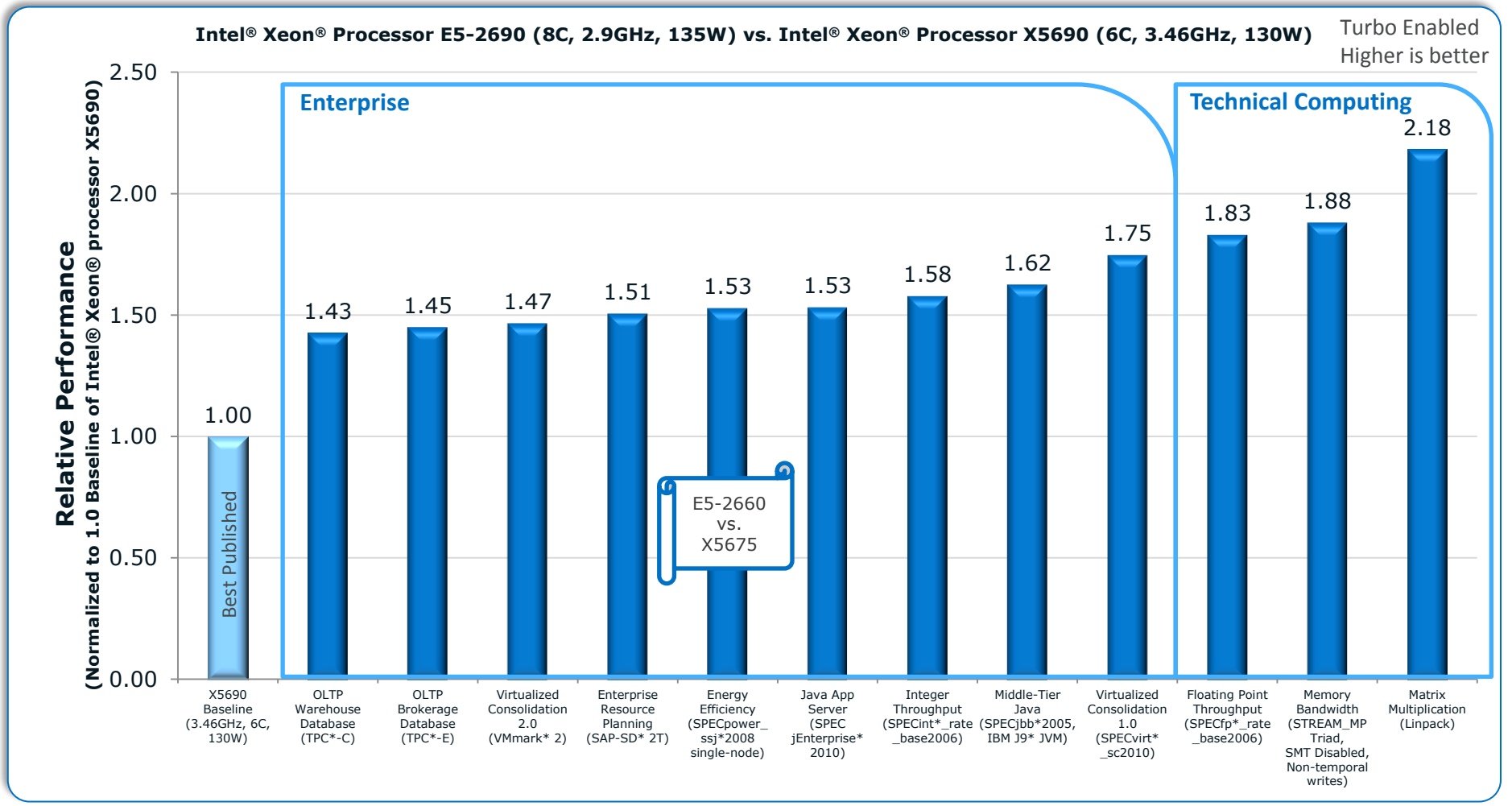
Intel® Xeon® processor E5-2690 delivers performance gains up to 2X

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Linpack performance may vary based on thermal solution.
Configuration Details: Please reference foil 24 for details.
For more information go to http://www.intel.com/performance

* Other names and brands may be claimed as the property of others

Intel® Xeon® Processor E5-2600 Product Family Generational Performance Summary



Intel® Xeon® processor E5-2690 delivers performance gains up to 2X

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

Linpack performance may vary based on thermal solution.
Source: Best published results as of 31 May 2012
Configuration Details: Please reference slide speaker notes.
For more information go to http://www.intel.com/performance

* Other names and brands may be claimed as the property of others

X-Gene™: 64-bit ARM CPU and SoC

Paramesh Gopi
Gaurav Singh
Greg Favor

8.29.2012

Cloud Computing Technology Trends

Fabric Interconnect between Rack Units



- **High Density Servers → " Sea of CPUs"**
 - Smaller & power-efficient CPUs; Beefier memory & IO subsystems
 - Distributed Fabric → networking & storage IO sharing & virtualization
- **Server-on-Chip (SoC) Approach**
 - Integrated NIC and IO chipset
 - CPU/ GPU combination for HPC applications
- **Active Power Management**
 - Firmware based optimization based on user Workload (Power is measured through TDP)
 - Maximize performance while managing TDP
- **Server Standardization**
 - Service provider specified
 - ODM designed & manufactured
 - Open Source/ non-commercial SW base
 - Open Stack, Open Compute

Cloud Servers - Typical Form Factors

Public Cloud

Applications

- Scale Out Services → Hosted Mail, Search, Social, Cloud Hosting

Platforms

- Dell PowerEdge C, HP ProLiant Microserver, DCS custom

Typical Specifications

- 1/2 Socket 2/4 core 2.8GHz, 80W
- 280 SpecIntRate
- System Power <500W; Cost <$2K





Building Out vs. Scaling Out



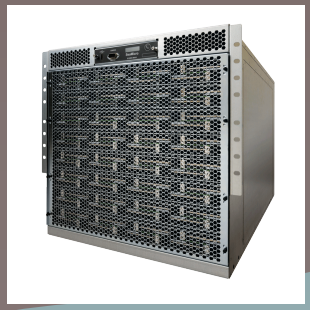
TODAY: 2 RU

- 2 Nodes per Rack Unit
- 2 Sockets @ 95W each
- Shared Chassis, Power Supply & Cooling
- Google, FB, Amazon Custom Datacenters

3 RU

- 8/12 Nodes in 3RU
- Single Socket @ 45W
- Shared Chassis, Power Supply, & Cooling
- Dell PowerEdge 5220, Supermicro MicroCloud





TOMORROW: 10 RU

- 256/ 512 Nodes in 10RU
- Single Socket @ 10-20W
- Shared IO Resources
- Integrated ToR Switch
- SeaMicro SM10000, HP Redstone

Opportunities from Hardware

Integration

- **Cores + memory + networking + I/O**
- Lower latency, better QoS
 - Multiple Priorities
 - B/W guarantees

Efficient Out-of Order Cores

- Break tradeoff between wimpy and brawny cores
- Energy efficiency at good performance (ARM-based processors are well suited here)

Virtualization Support

- Improve utilization without hurting performance

Highly Integrated Server on Chip

Efficient Low Latency Interconnect

Cloud Requirements →
Integrated, Right-Sized Compute. Memory. Network.

ARMv8 (Oban): Fully Backwards Compatible
New 64b ISA + Current 32b ISA

New: ARMv8

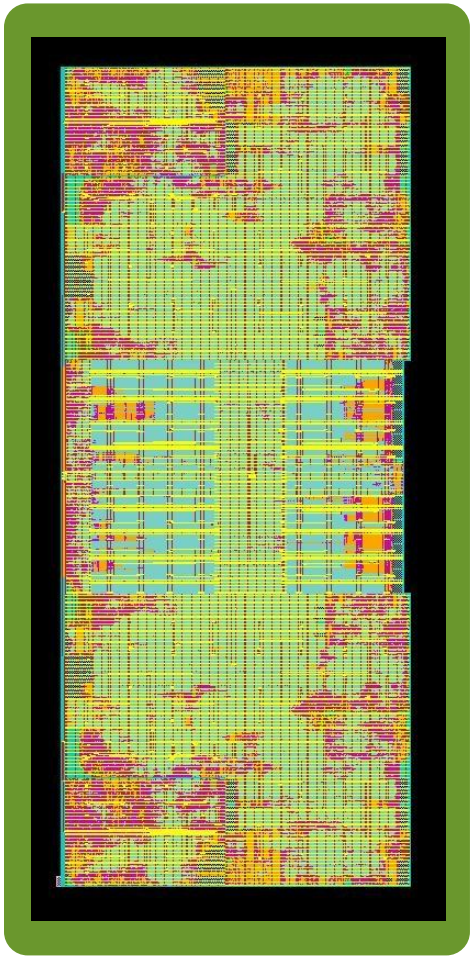
- 64b (General) and 128b (FP,SIMD) registers
- SP, PC no longer general purpose registers
- Uniform load/store addressing modes
- Larger data and instr. offset ranges
- Simplified load/store multiple instructions

- Reduced conditional instructions
- 32 128b FP/SIMD architecture registers
- No SIMD on general purpose registers
- New instructions for debug, TLB, barriers
- New Crypto acceleration instructions

ARMv8

- **New High Performance 64bit ISA + compatibility with existing 32bit ISA**
- **Full CPU, IO, Interrupt, Timer Virtualization**
- **Enhanced 128b SIMD operations**
- **High performance Floating-Point operations including FMADD**
- **Standard Performance Monitoring, Instr. Trace and Debug Architecture**

X-Gene™ CPU Design Goals



- **High-Performance Low-Power Microarchitecture**
 - Design point targets balance between performance, power, and size
 - Maximum “bang for the buck”
- **Low Power Microarchitecture Features**
 - Sophisticated branch prediction, Caches, Unified register renaming
 - Minimal instruction replay cases
 - Separate smaller schedulers per pipe
 - Full set of power management features
- ***Good*** **Single-Thread Performance, but also Efficiently Scalable to Many Cores**
 - Scalable CPU and interconnect architecture 2-128 cores
 - High bandwidth, low latency switch fabric > 1Tbps
 - High-performance distributed hardware cache coherency
- **Technology Portability**
 - Fully synthesizable RTL
 - Semi-custom cell-based design methodology
 - Small targeted set of custom macros (plus clock distribution cells/macros)

X-Gene™ Processor Module

Processor Module

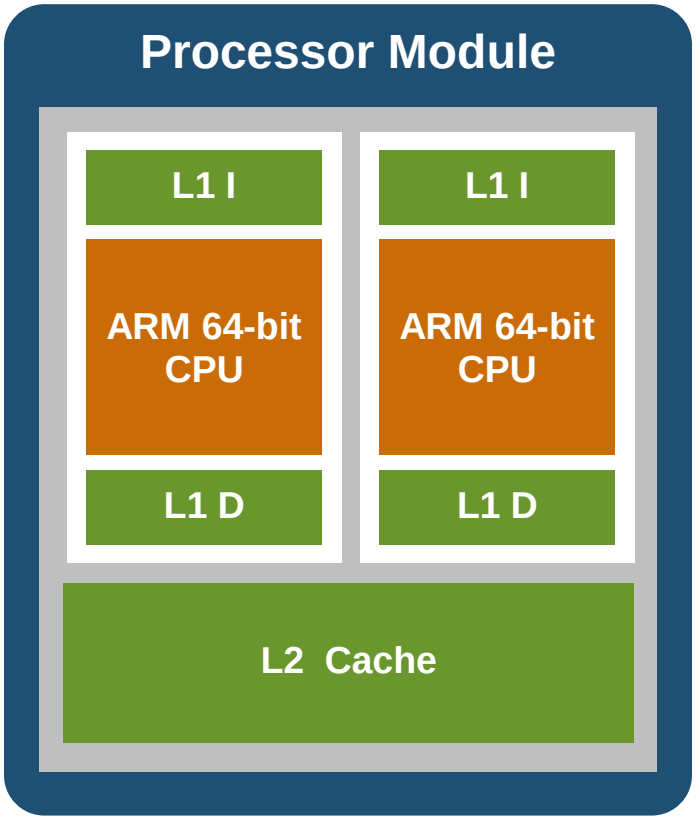
- 2 cores + shared L2 cache
- 4 wide out-of-order superscalar microarchitecture
- Integer, scalar, HP/SP/DP FPU and 128b SIMD engine
- Hardware virtualization support
- Hardware tablewalk and nested page tables
- Full set of static and dynamic power management features
 - Fine grain/macro clock gating, DVFS
 - C0, C1, C3, C4, C6 states

Cache Hierarchy

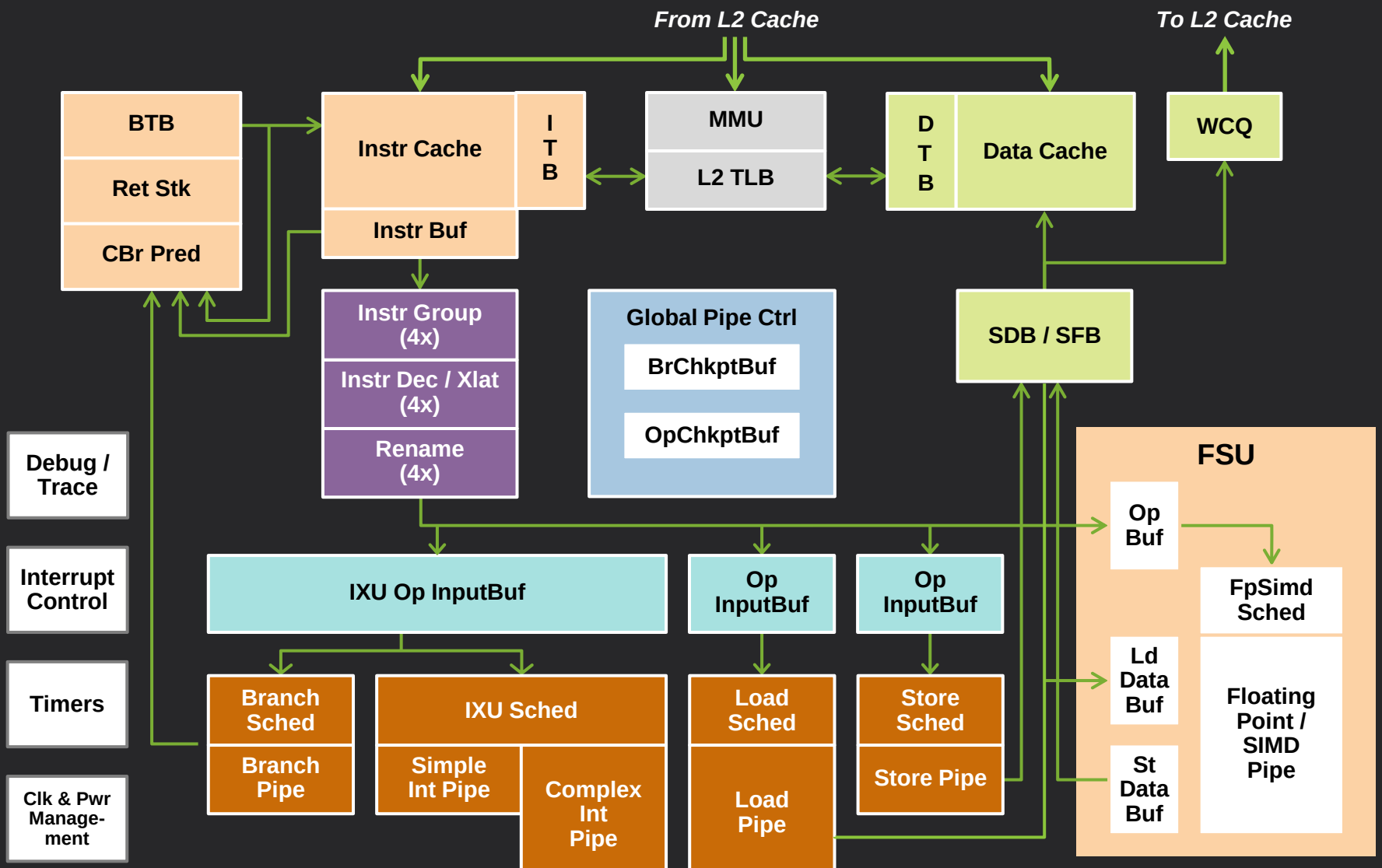
- Separate L1I and L1D caches
- Shared L2 cache among 2 CPUs
- Last-level globally shared L3 Cache
- Advanced hardware prefetch in L1 and L2
- L2 inclusive of L1 write-thru data caches

RAS

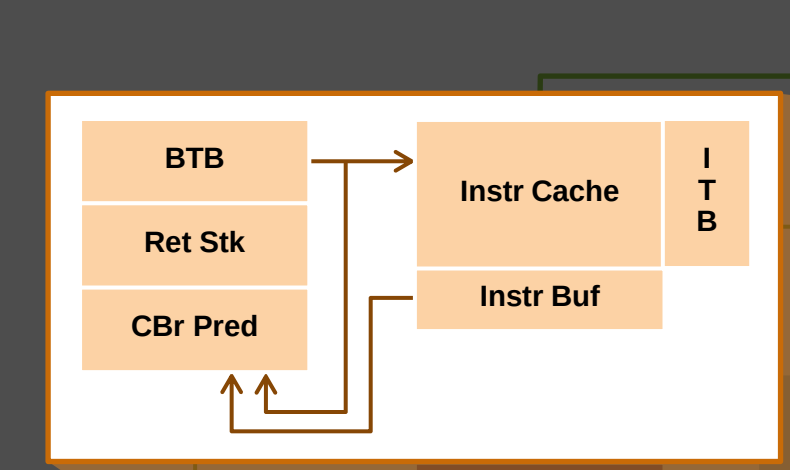
- ECC and Parity protection of all Caches,Tags,TLBs
- Data poisoning and error isolation



X-Gene™ CPU Block Diagram



X-Gene™ Instruction Fetch



From L2 Cache

To L2 Cache

MMU

D
T

Data Cache

WCQ

Instr Dec / Xlat (4x)

Rename (4x)

Debug / Trace

Interrupt Control

Timers

Clk & Pwr Manage-ment

IXU Op InputBuf

Branch Sched

Branch Pipe

IXU Sched

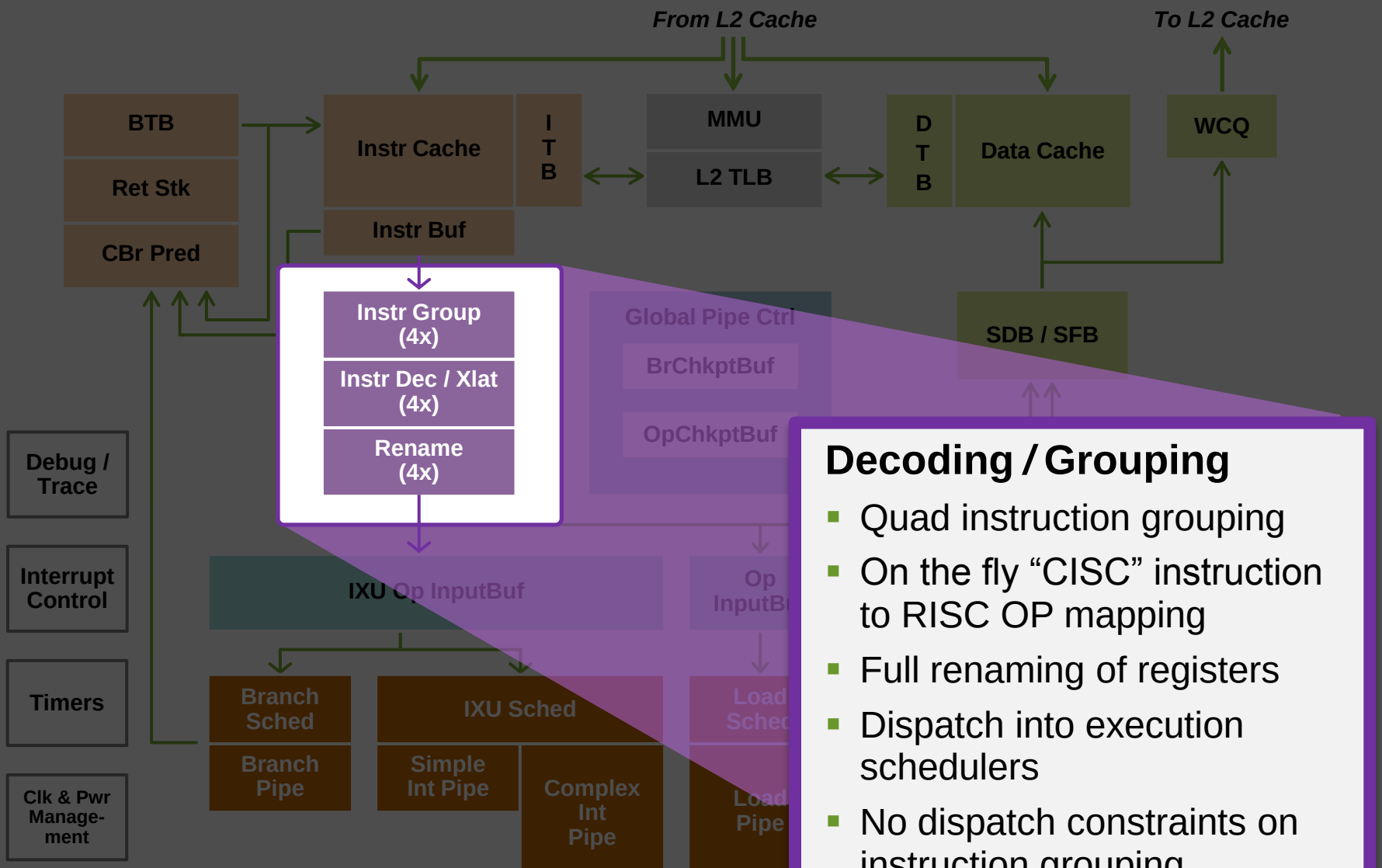
Simple Int Pipe

Complex Int Pipe

I-Cache & Fetch Unit

- Fetch multiple instructions / cycle
- Instruction pre-decode bits stored with each cache-line
- Single cycle scan to pick next predicted taken branch
- 2-level branch prediction
- Branch Target Buffer
- Conditional, call/return branch predictors
- History based indirect branch predictors
- First level fully-associative L1 TLB

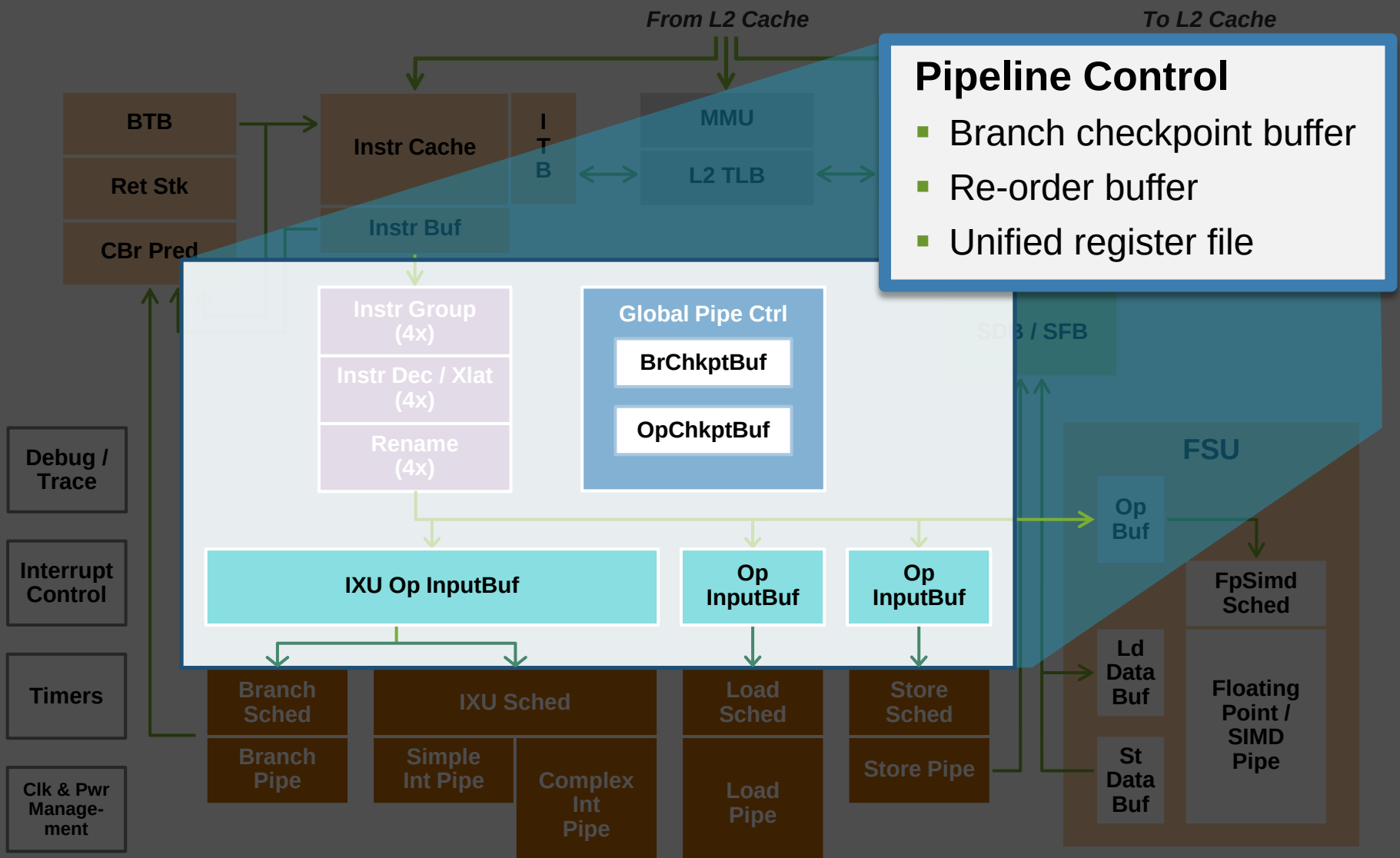
X-Gene™ Instructions Decoding/Grouping



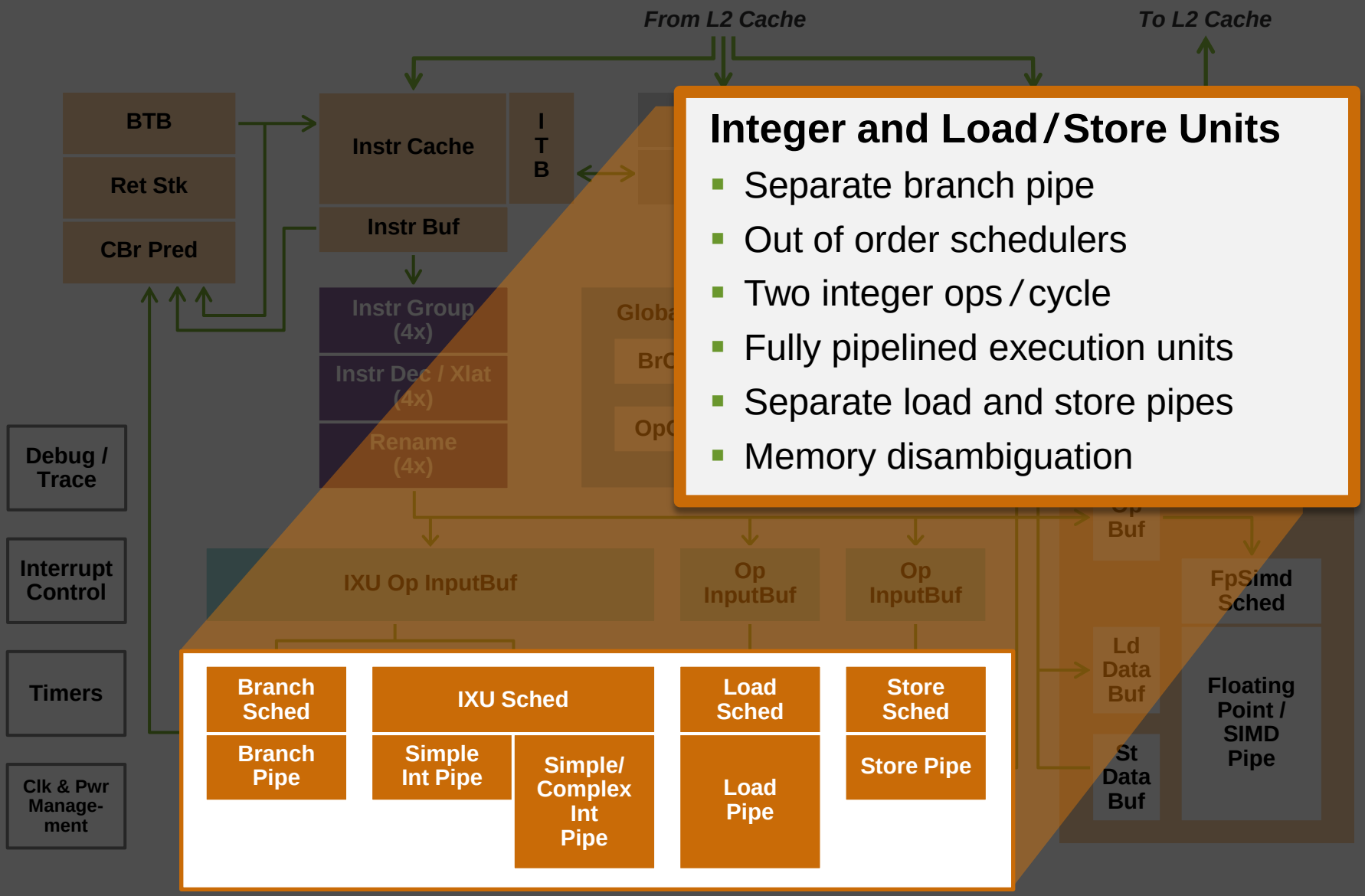
Decoding / Grouping

- Quad instruction grouping
- On the fly “CISC” instruction to RISC OP mapping
- Full renaming of registers
- Dispatch into execution schedulers
- No dispatch constraints on instruction grouping

X-Gene™ Reorder Buffer, Dispatch and Control



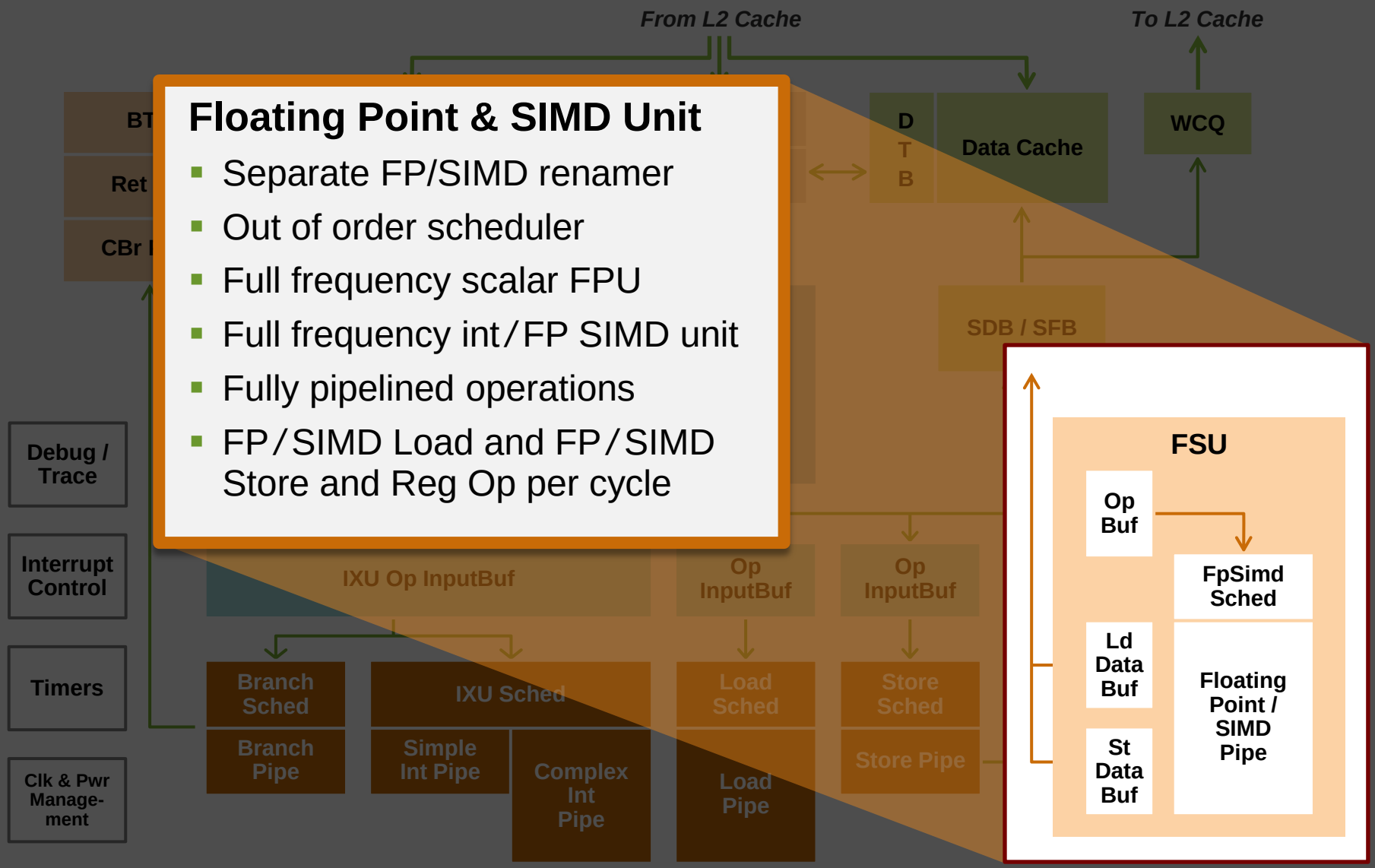
X-Gene™ Integer, Branch, Load and Store Units



Integer and Load / Store Units

- Separate branch pipe
- Out of order schedulers
- Two integer ops / cycle
- Fully pipelined execution units
- Separate load and store pipes
- Memory disambiguation

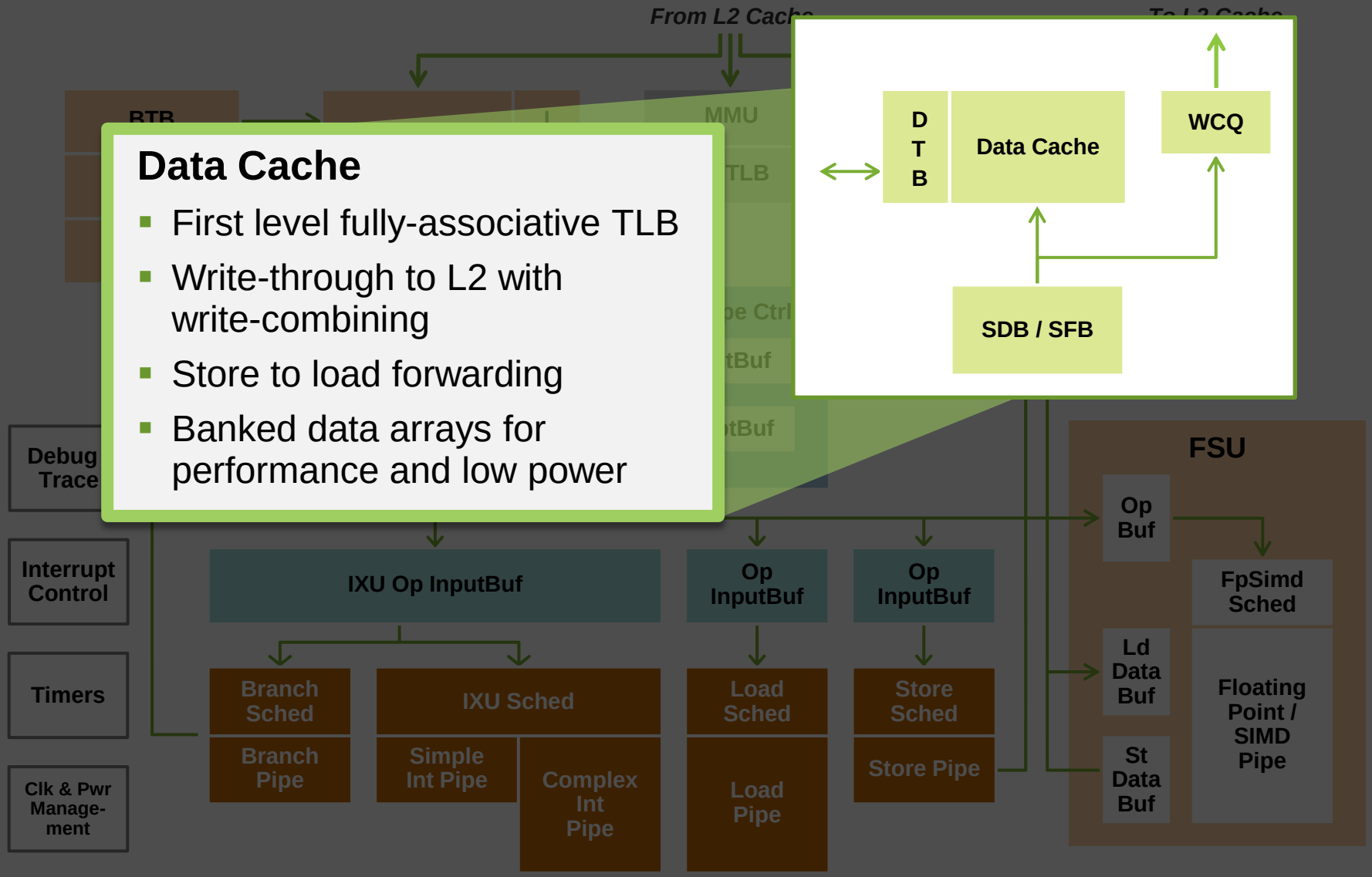
X-Gene™ FPU/SIMD



Floating Point & SIMD Unit

- Separate FP/SIMD renamer
- Out of order scheduler
- Full frequency scalar FPU
- Full frequency int/FP SIMD unit
- Fully pipelined operations
- FP/SIMD Load and FP/SIMD Store and Reg Op per cycle

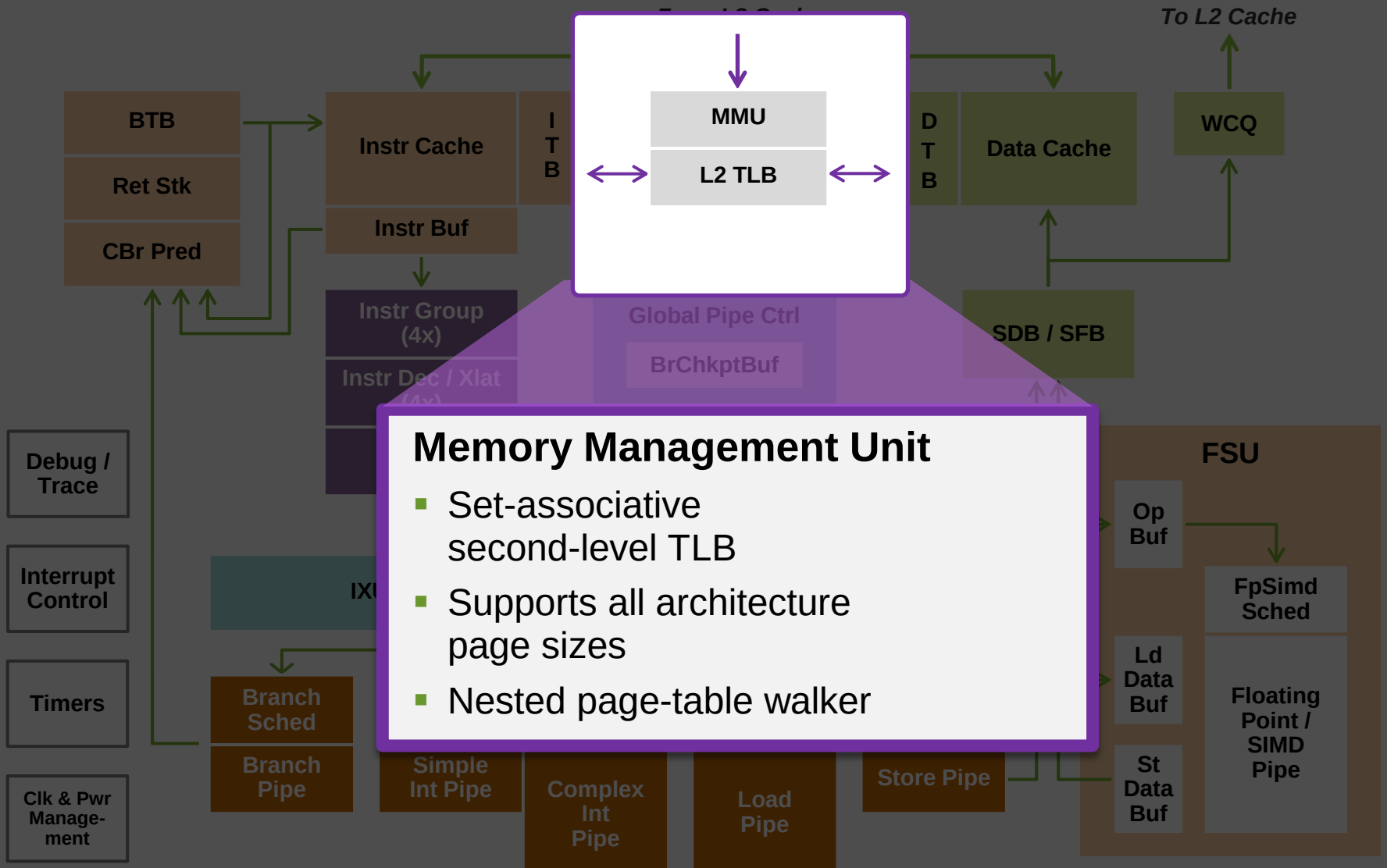
X-Gene™ Data Cache



Data Cache

- First level fully-associative TLB
- Write-through to L2 with write-combining
- Store to load forwarding
- Banked data arrays for performance and low power

X-Gene™ Memory Management



Memory Management Unit

- Set-associative second-level TLB
- Supports all architecture page sizes
- Nested page-table walker

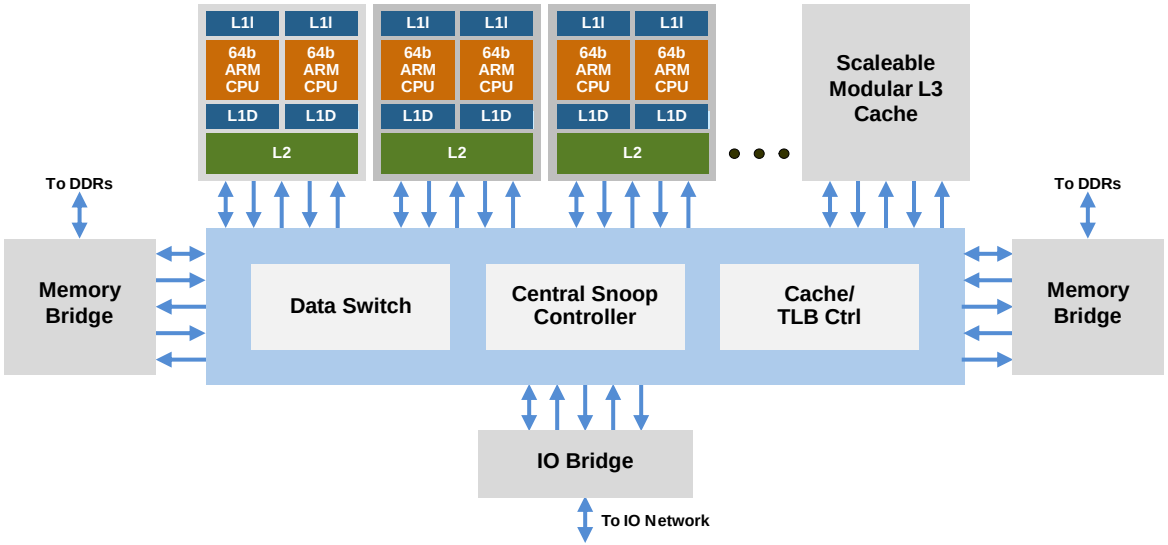
X-Gene™ – CPU Memory Subsystem

High-performance Symmetric Multi-core Design

- Modular architecture
- Three level cache hierarchy
- Globally shared L3 Cache

Coherent Network

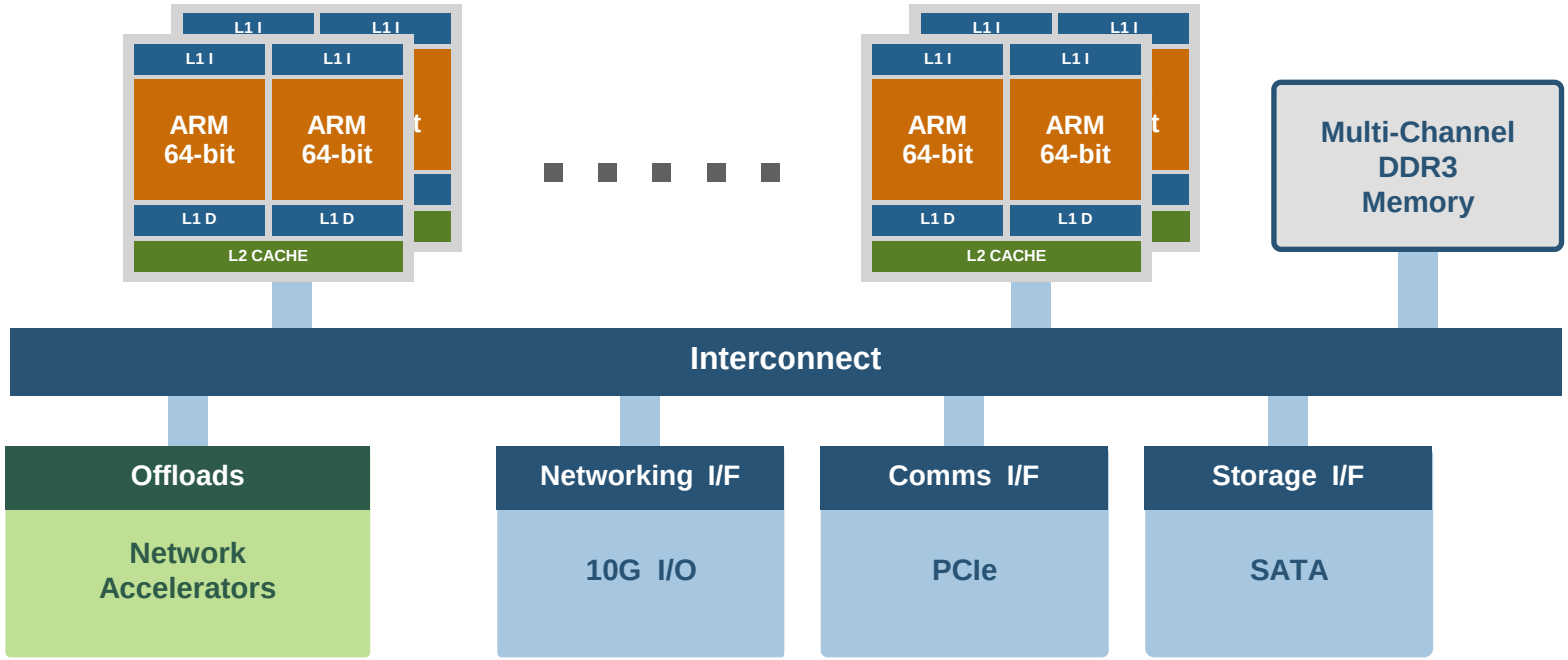
- Runs at full CPU frequency
- <15ns latency, ~200GB/s B/W
- Over 400 transactions in flight
- Central snoop controller and ordering point
- Decoupled frequency and power domains
- Support global cache and TLB inv operations



Bridges

- Memory Bridges to DRAM interfaces
- IO Bridge for SOC connectivity

X-Gene™ Server on Chip



64bit ARM Server Class CPU → Multi-core for Distributed Computing

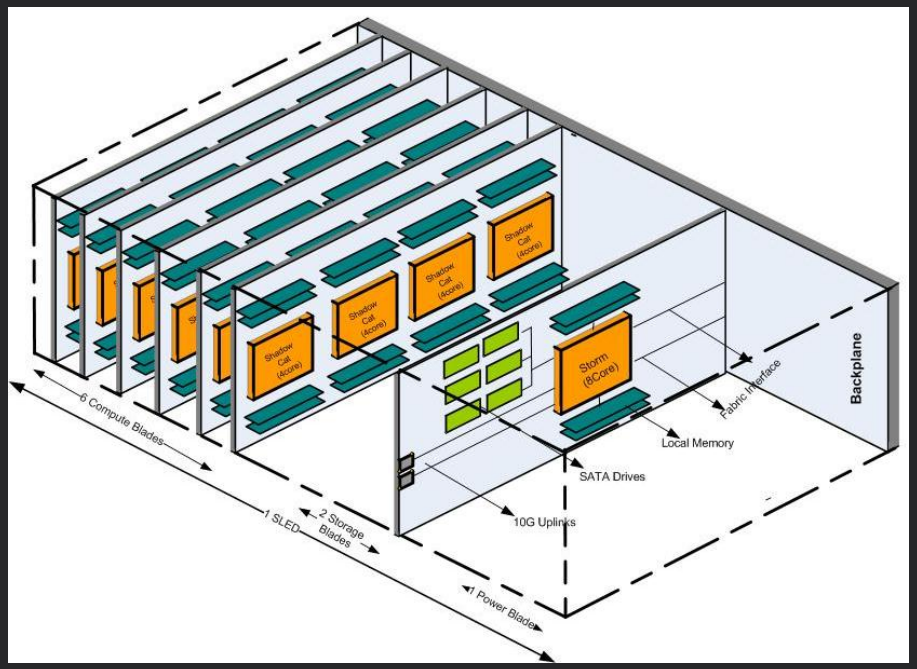
Increased Memory Capacity and 10G I/O Integration

Integrated Peripherals and L2 Switching

Workload Specific Acceleration

Available: 2H'12

Right Sizing + Connected On-Chip Fabric



Customizable Blade Design

- Configurable swappable blades within 1 sled
- Networking/ Compute/ Storage shared over common bed of CPUs → Saves Power & Cost

Overall System Optimization

- Integrated NIC and IO chipset
- Load Balancing Across multiple blades to Optimize System Balance
- Shared Resources for System Management, Power and Cooling

System Capabilities

- 1000s of CPU cores in 10RU
- 100s of CPU cores per blade
- 100s of Gbps of network bandwidth
- 10s of Tbps of interconnect fabric bandwidth

X-Gene™

Worlds First True Low Power Server-on-Chip™

A Perfect Storm Cometh …

SPARC64™ X:
Fujitsu's New Generation 16 Core Processor for the next generation UNIX servers

August 29, 2012

Takumi Maruyama

Processor Development Division
Enterprise Server Business Unit
Fujitsu Limited

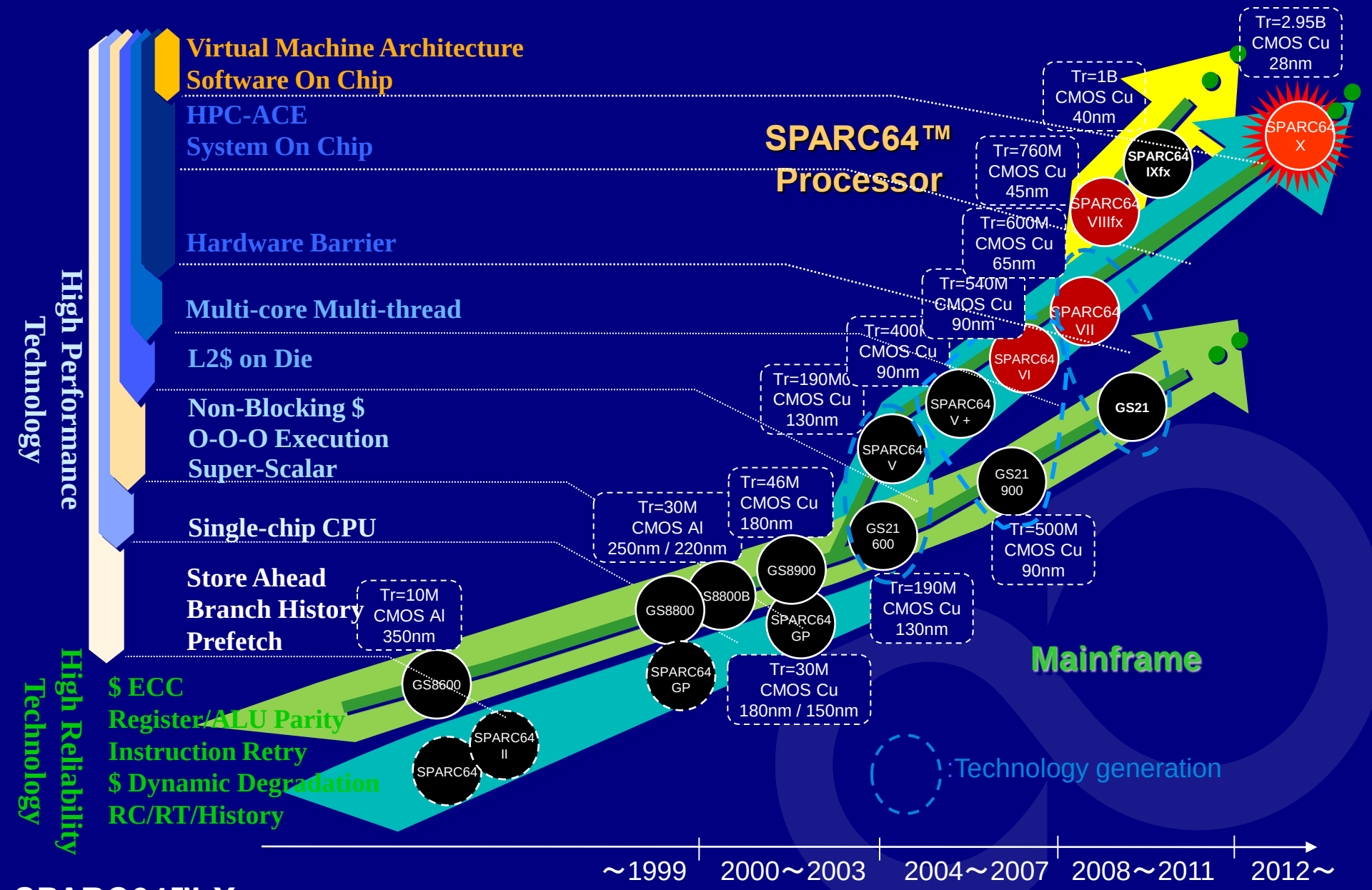
All Rights Reserved,Copyright© FUJITSU LIMITED 2012

Agenda

- Fujitsu Processor Development History
- SPARC64™ X
 - Design concept
 - SWoC (Software on Chip)
 - Processor chip overview
 - u-Architecture
 - Performance
- Summary

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

Fujitsu Processor Development



All Rights Reserved, Copyright© FUJITSU LIMITED 2012

SPARC64™ X Design Concept

- Combine UNIX and HPC FJ processor features to realize an extremely high throughput UNIX processor.
 - SPARC64 VII/VII+ (UNIX processor) feature
 - High CPU frequency (up-to 3GHz)
 - Multicore/Multithread
 - Scalability : up-to 64sockets
 - SPARC64 VIIIfx (HPC processor) feature
 - HPC-ACE: Innovative ISA extensions to SPARC-V9
 - High Memory B/W: peak 64GB/s, Embedded Memory Controller
- Add new features vital to current and future UNIX servers
 - Virtual Machine Architecture
 - Software On Chip
 - Embedded IOC (PCI-GEN3 controller)
 - Direct CPU-CPU interconnect

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

Software on Chip 1/2

- HW for SW
 Accelerates specific software function with HW

- The targets
 - Decimal operation (IEEE754 decimal and NUMBER)
 - Cypher operation (AES/DES)
 - Database acceleration

- HW implementation
 - The HW engines for SWoC are implemented in FPU
 - To fully utilize 128 FP registers & software pipelining
 - Implemented as instructions rather than dedicated co-processor to maximize flexibility of SW.
 - Avoid complication due to “CISC” type instructions
 - Various “RISC” type instructions are newly defined, instead.
 - 18 insts. for Decimal, and 10 insts. for Cypher operation

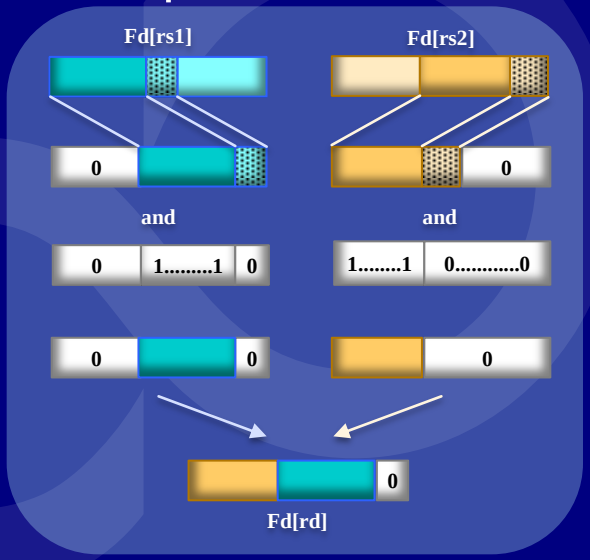
All Rights Reserved, Copyright© FUJITSU LIMITED 2012

Software on Chip 2/2

Decimal Instructions

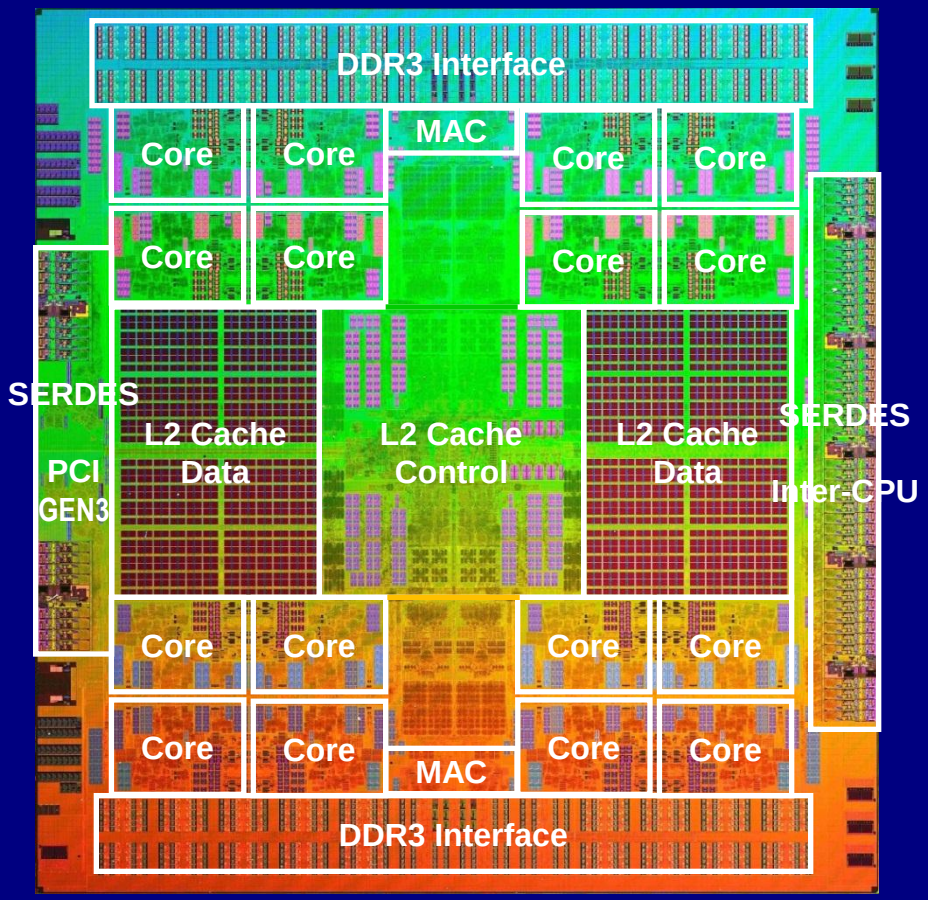
- Supported data type
 - IEEE754 DPD(Densely Packed Decimal)
 8B fixed length
 - NUMBER
 Variable length (max 21Byte)

- Instructions
 - Both DPD/NUMBER instructions are defined as 8B operation (add/sub/mul/div/cmp) on FP registers
 - To maximize performance with reasonable HW cost
 - When the data length is > 8byte, multiple such instructions will be used.
 - An instruction for special byte-shift on FP registers is newly added to support unaligned NUMBER



All Rights Reserved, Copyright© FUJITSU LIMITED 2012

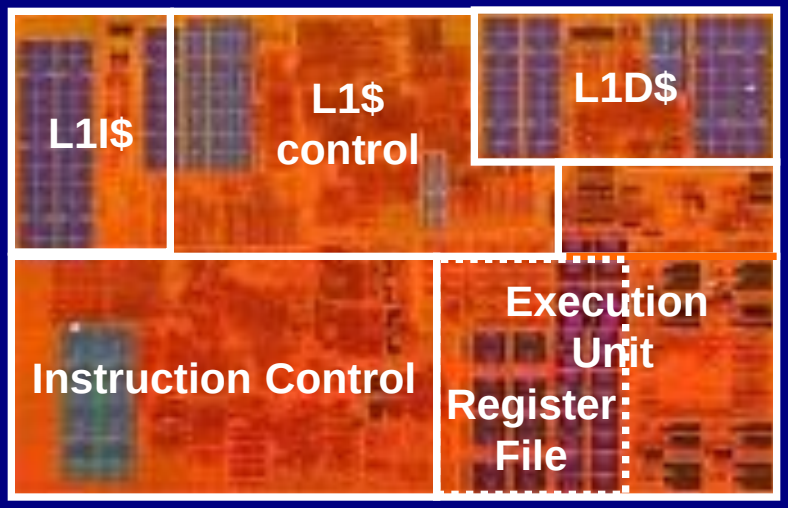
SPARC64™ X Chip Overview



- **Architecture Features**
 - 16 cores x 2 threads
 - SWoC (Software on Chip)
 - Shared 24 MB L2$
 - Embedded Memory and IO Controller
- **28nm CMOS**
 - 23.5mm x 25.0mm
 - 2,950M transistors
 - 1,500 signal pins
 - 3GHz
- **Performance (peak)**
 - 288GIPS/382GFlops
 - 102GB/s memory throughput

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

SPARC64™ X Core spec



Instruction Set Architecture	SPARC-V9/JPS HPC-ACE VM SWoC
Branch Prediction	4K BRHIS 16K PHT
Integer Execution Units	156 GPR x 2 + 64 GUB
	ALU/SHIFT x2 ALU/AGEN x2 MULT/DIVIDE x1
FP Execution Units	128 FPR x 2 + 64 FUB
	FMA x4, FDIV x2 IMA/Logic x4 Decimal x1 / Cypher x2
L1$	L1I$ 64KB/4way L1D$ 64KB/4way

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

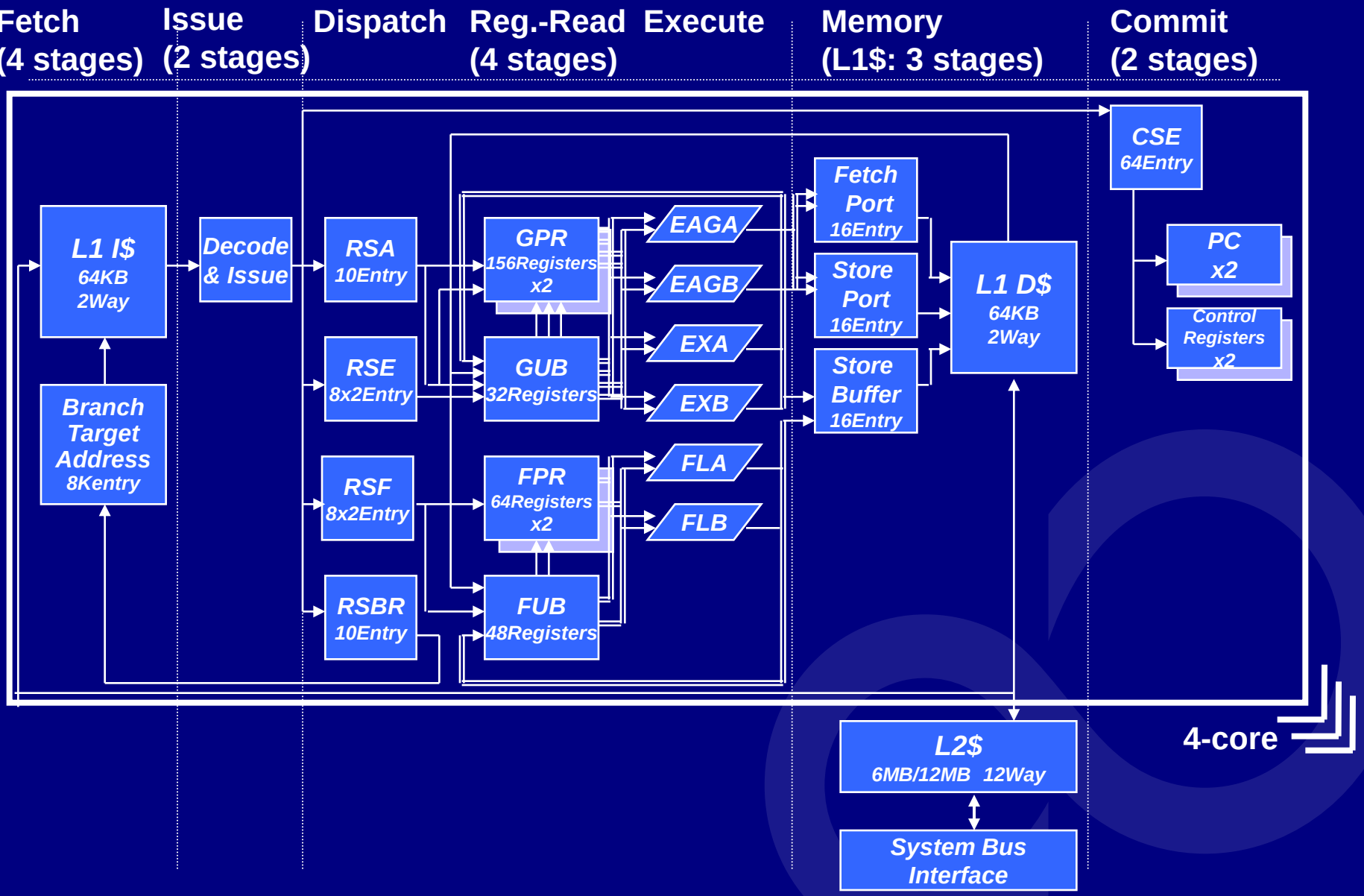
u-Architecture enhancements from SPARC64™ VII+

- CPU Core
 - Deeper pipeline to increase Frequency
 - Better Branch Prediction Scheme
 - Various Queue-size and #Floating point register increase
 - Richer execution Units, including
 - 2EX + 2EAG → 2EX + 2EX/EAG
 - 2FMA → 4FMA to support 2way-SIMD
 - SWoC engine (Decimal and Cypher)
 - More aggressive O-O-O execution of load and store
 - Multi-banked 2port L1-Cache

- System On Chip
 - #core and L2$ size (4core/12MB→16core/24MB)
 - Memory Controller, IO Controller, and CPU-CPU I/F are all embedded to increase performance and reduce cost.

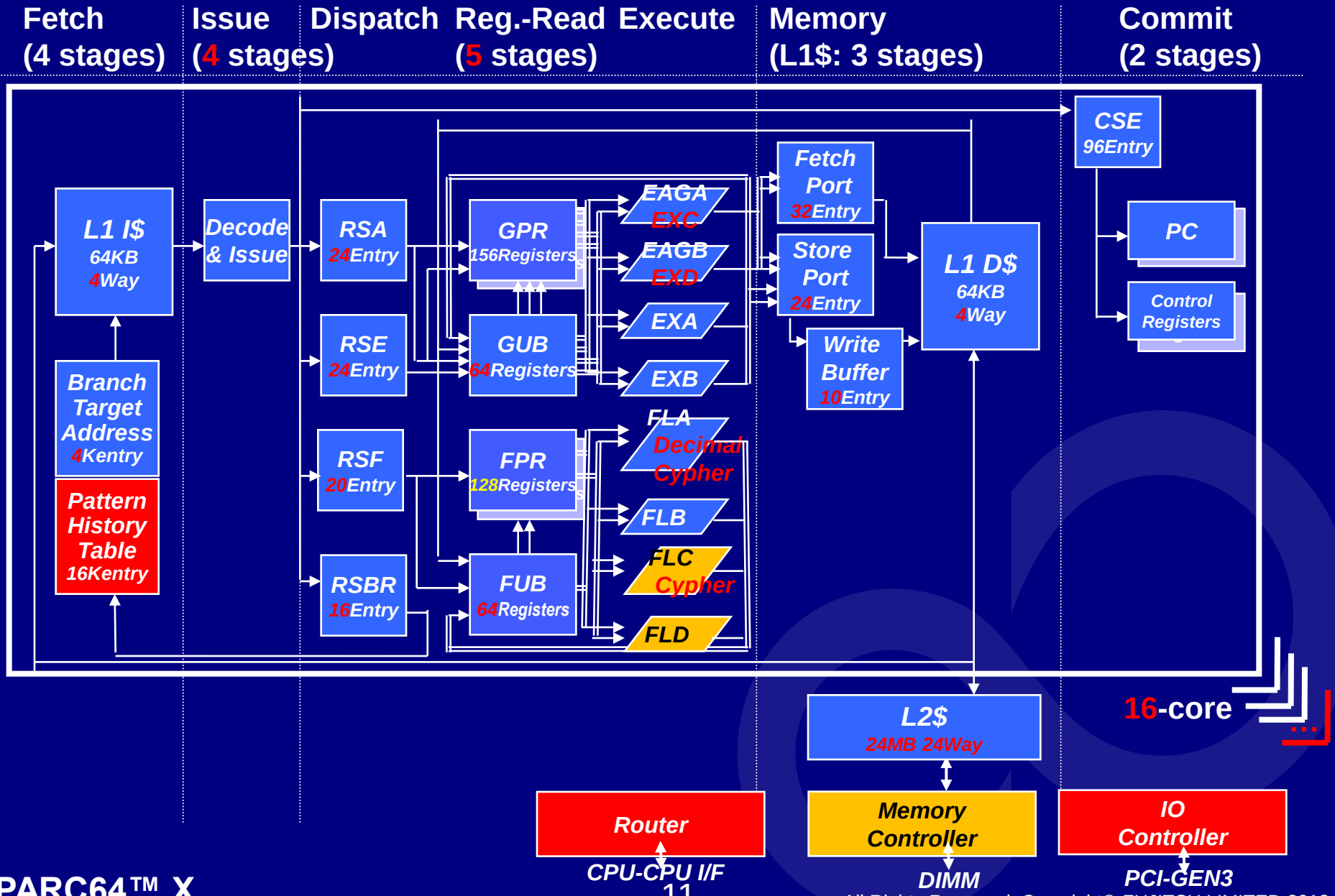
All Rights Reserved, Copyright© FUJITSU LIMITED 2012

SPARC64™ VII/VII+ Pipeline



All Rights Reserved, Copyright© FUJITSU LIMITED 2012

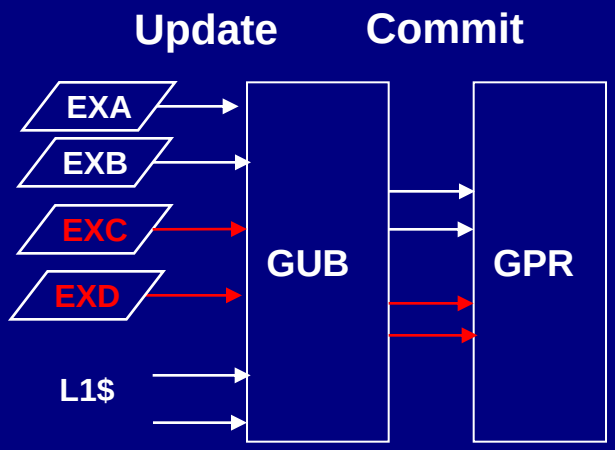
SPARC64™ X Pipeline



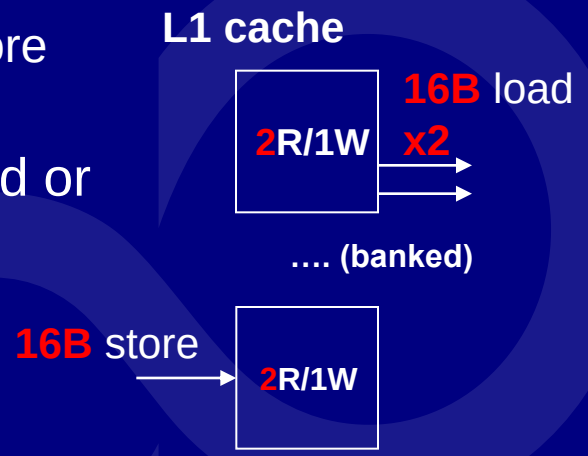
All Rights Reserved, Copyright© FUJITSU LIMITED 2012

Execution units enhancements (Ex.)

- **Integer Execution Unit**
 - 2EX + 2EAG → 2EX + 2EX/EAG
 - 2 → 4W GPR
 - → 4 integer instructions can be executed per cycle (sustained)



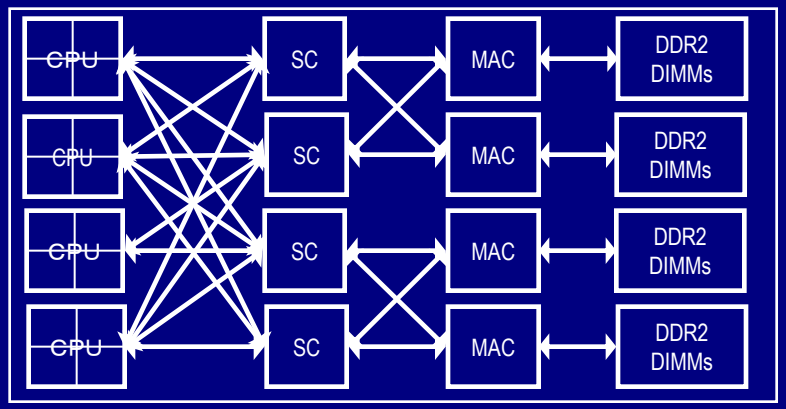
- **Load Store Unit**
 - Aggressive load/store O-O-O execution:
 - Execute load without waiting for preceding store address calculation.
 - Multi-banked 2port L1-cache to execute 2 load or 1 load+1 store in parallel
 - Doubled L1$ bus size
 - Doubled L1$ associativity (2→4way)
 - → Increase L1-cache throughput and hit-rate



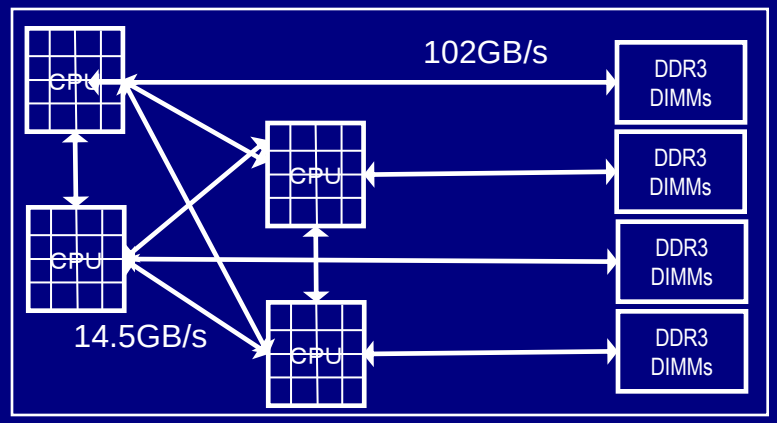
All Rights Reserved, Copyright© FUJITSU LIMITED 2012

SPARC64™ X interconnects

SPARC64™ VII/VII+ interconnects
(SPARC Enterprise M8000)



SPARC64™ X interconnects



- **SPARC64™ VII/VII+ interconnects**
 - 4 CPU require 8 additional LSIs to be connected with DIMM
 - DIMM i/f: 4.35GB/s (STREAMtriad)

- **SPARC64™ X interconnects**
 - No additional LSIs to be connected with DIMM
 - DIMM i/f: 65.6GB/s (STREAMtriad)
 - CPU i/f: 14.5GB/s x 5ports (peak)
 - 3 ports: glueless 4way CPU interconnect
 - 2 ports: > 4way CPU

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

High Speed Transceivers (SerDes)

- CPU-CPU glue-less communication links
 - 14.5Gb/s x 8 lanes bi-directional serial interface, 5 ports
 - Embedded equalizer circuit enables long distance signal transmission
 - Embedded adaptive control logic optimizes equalizer parameters automatically depending on the various system configurations
- PCI Express ports
 - 8Gb/s x 8 lanes (Gen 3), 2 ports



14.5Gb/s x 8lanes SerDes

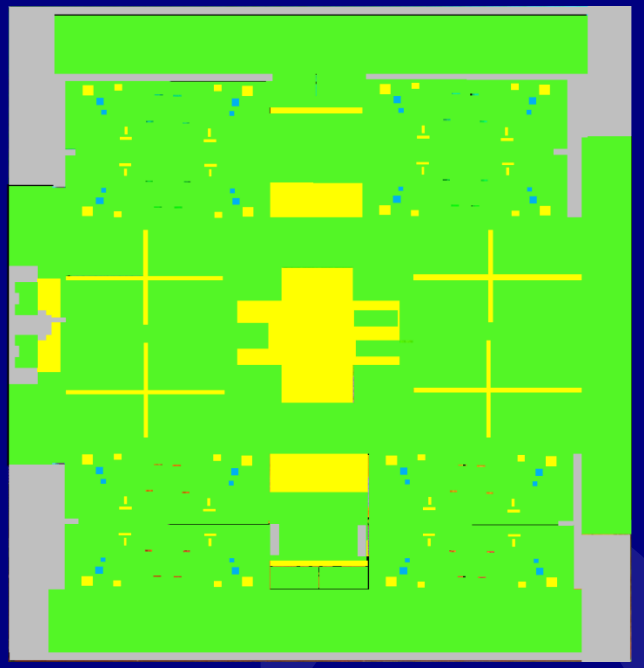
- Built-in SerDes provides peak 88.5GB/s x2 (up/down) total throughput

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

Reliability, Availability, Serviceability

Units	Error detection and correction scheme
Cache (Tag)	ECC Duplicate & Parity
Cache (Data)	ECC Parity
Register	ECC (INT/FP) Parity(Others)
ALU	Parity/Residue
Cache dynamic degradation	Yes
HW Instruction Retry	Yes
History	Yes

SPARC64™ X RAS diagram



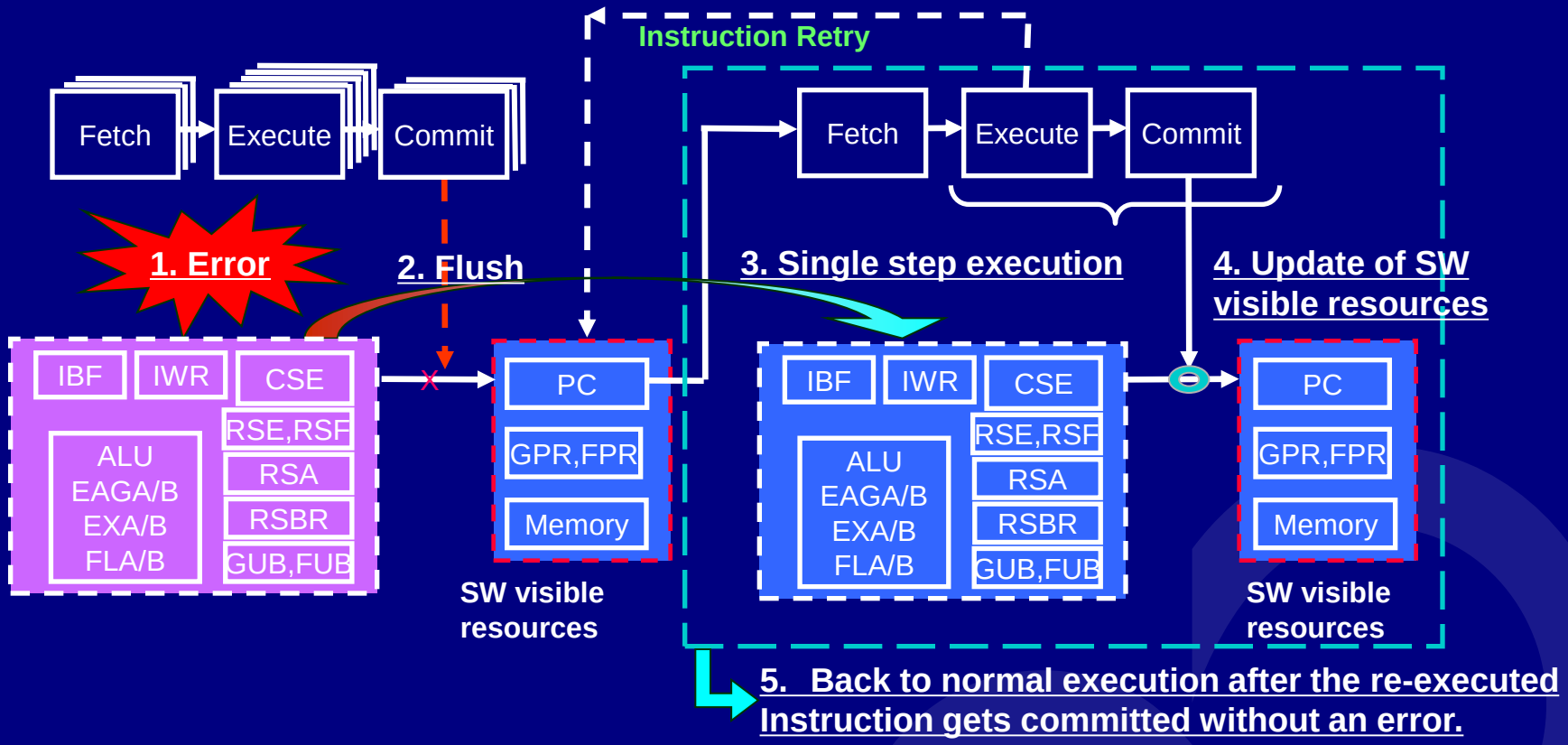
Green: 1bit error Correctable
Yellow: 1bit error Detectable
Gray: 1bit error harmless

- New RAS features from SPARC64™ VII/VII+
 - Floating-point registers are ECC protected
 - #checkers increased to ~53,000 to identify a failure point more precisely

→ Guarantees Data Integrity

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

Hardware Instruction Retry

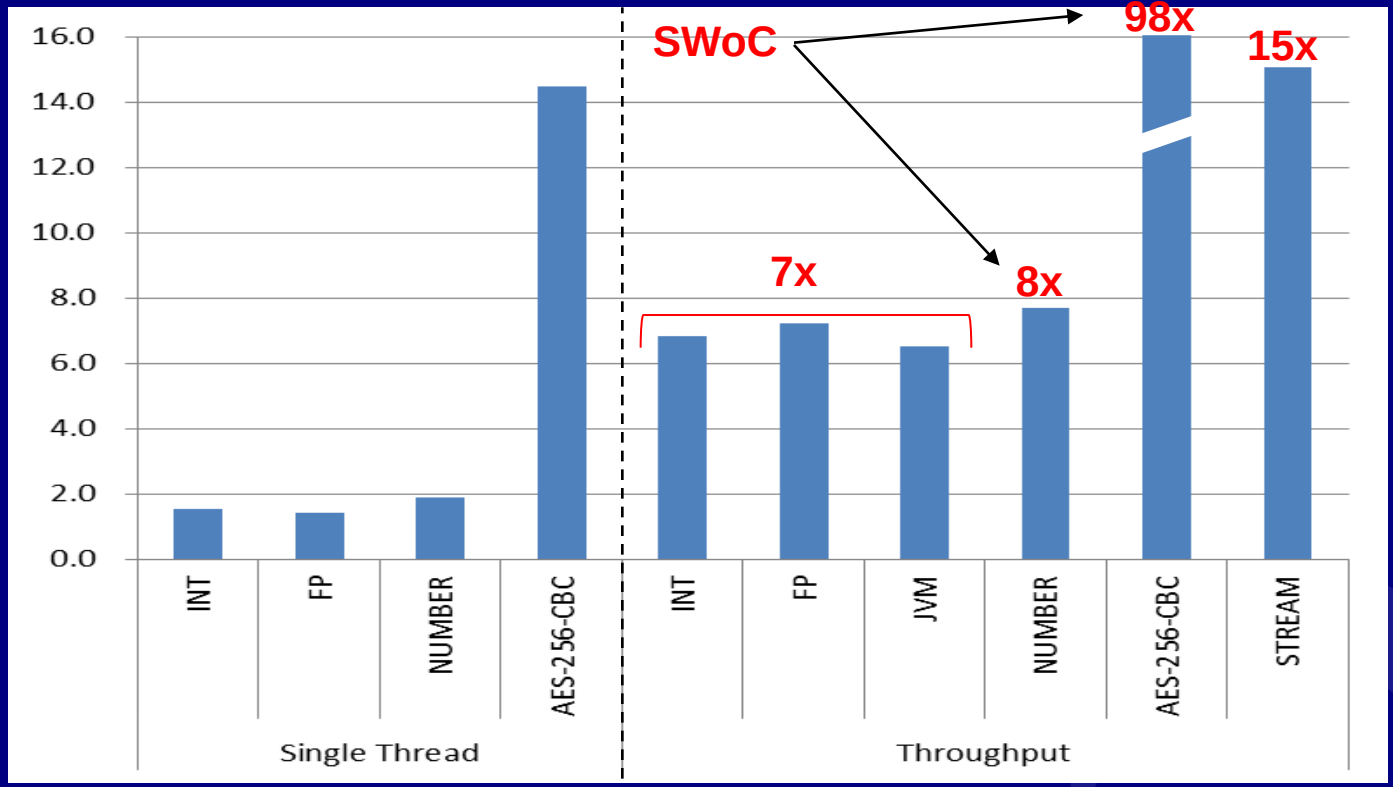


- When an error is detected, Hardware re-execute the instruction automatically to remove the transient error by itself.

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

SPARC64™ X Performance @3GHz

Hardware measured results



- SPARC64™ X realizes 7x INT/FP/JVM throughput and 15x memory throughput of SPARC64™ VII+
 - The INT/FP/JVM result is with un-tuned Compiler/JVM.
- SWoC of SPARC64™ X results in max 98x throughput.
 - The NUMBER score is for scalar. Expect to be much better for vector data.

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

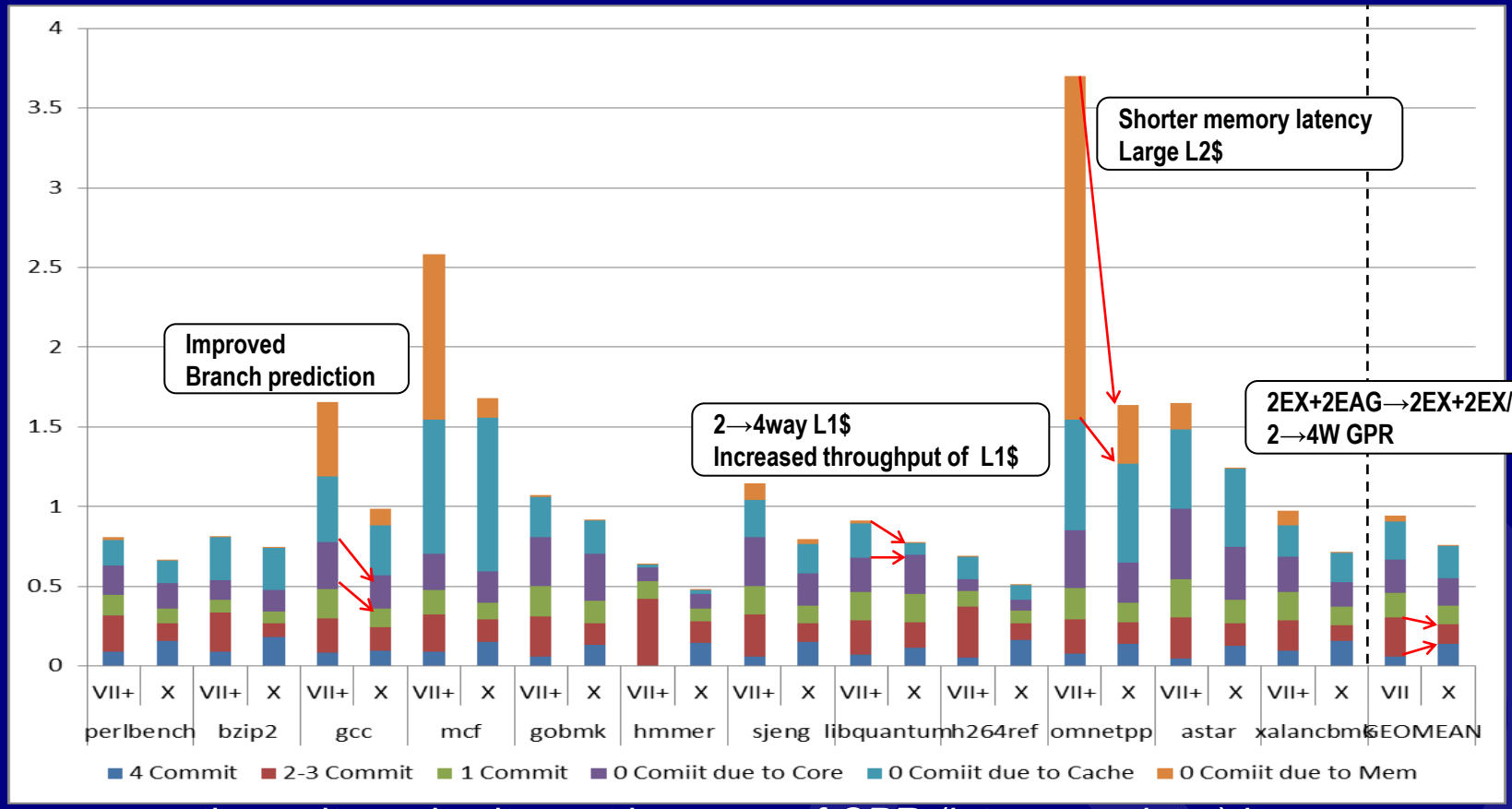
SPARC64™ X CPI (Cycle Per Instruction) Example

SPARC64™ VII+ v.s. SPARC64™ X
INT (single thread)

Hardware measured results

Lower Performance

Higher Performance



- 4 integer execution units and write port increase of GPR (integer register) improves overall performance.
- Memory latency reduction, Large L2$, branch prediction, and L1$ improvement also contribute to the high performance dramatically.

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

Summary

- SPARC64™ X is Fujitsu's10th SPARC processor which has been designed to be used for Fujitsu's next generation UNIX server.

- SPARC64™ X integrates 16cores + 24MB L2 cache with over 100GB/s(peak) memory B/W.
- SPARC64™ X keeps strong RAS features.

- SPARC64™ X chip is up and running in the lab.
- It has shown 7 times throughput of SPARC64™ VII+ w/o compiler tuning.
- SWoC is effective to accelerate specific software functions

- Fujitsu will continue to develop SPARC64™ series.

All Rights Reserved, Copyright© FUJITSU LIMITED 2012

Abbreviations

- SPARC64™ X
 - IB: Instruction Buffer
 - RSA: Reservation Station for Address generation
 - RSE: Reservation Station for Execution
 - RSF: Reservation Station for Floating-point
 - RSBR: Reservation Station for Branch
 - GUB: General Update Buffer
 - FUB: Floating point Update Buffer
 - GPR: General Purpose Register
 - FPR: Floating Point Register
 - CSE: Commit Stack Entry

All Rights Reserved, Copyright© FUJITSU LIMITED 2012


ORACLE®



ORACLE®

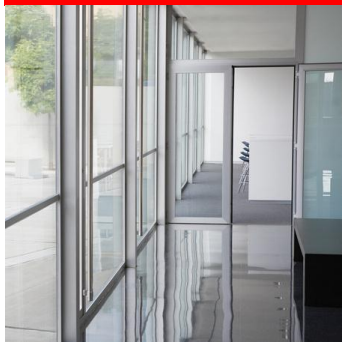
SPARC T5: 16-core CMT Processor with Glueless 1-Hop Scaling to 8-Sockets

Sebastian Turullols and Ram Sivaramakrishnan
Hardware Directors, Microelectronics

HOT CHIPS 24

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle’s products remains at the sole discretion of Oracle.

Outline



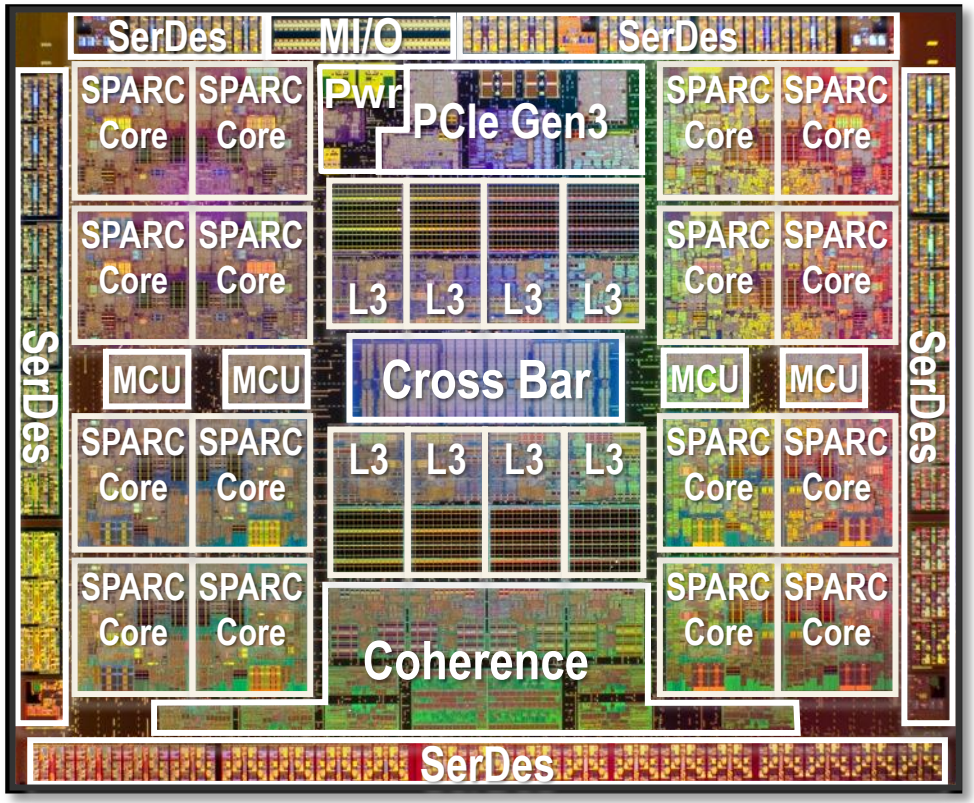
- Design Objectives
- SPARC T5 Processor Overview
- Core S3
- Cache Hierarchy Components
- Internode Coherency for 8-Socket Scaling
- Power Management Advances
- PCI-Express Gen3 I/O Subsystem
- Summary

SPARC T5 Design Objectives

- Multiply performance
- Achieve highly efficient 8-socket glueless 1-hop scalability
- Optimize for Oracle workloads and Engineered Systems
- Maximize power efficiency
- Provide Enterprise Class RAS



T5 Processor Overview



- 16 S3 cores @ 3.6GHz
- 8MB shared L3 Cache
- 8 DDR3 BL8 Schedulers providing 80 GB/s BW
- 8-way 1-hop glueless scalability
- Integrated 2x8 PCIe Gen 3
- Advanced Power Management with DVFS

S3 Core Recap

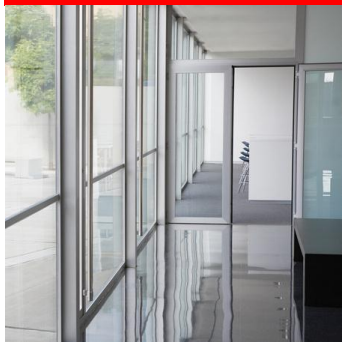
- 28nm port from 40nm T4
- Out-of-order, dual-issue
- High frequency achieved with 3.6GHz
 16 stage integer pipe
- Dynamically threaded, one to eight strands
- Accelerates 16 encryption algorithms and
 random number generation

SPARC T5 Leads in On-Chip Encryption Acceleration

- Built in, zero-overhead crypto
- Works with Solaris ZFS file system for faster file system encryption
- Provides secure consolidation with dynamic VM migration

On-Chip Accelerators	SPARC T5	IBM Power7	Intel Westmere/ Sandybridge
Asymmetric /Public Key Encryption	RSA, DH, DSA, ECC	none	RSA, ECC
Symmetric Key / Bulk Encryption	AES, DES, 3DES, Camellia, Kasumi	none	AES
Message Digest / Hash Functions	CRC32c, MD5, Sha-1, SHA-224, SHA-256, SHA-384, SHA-512	none	none
Random Number Generation	Supported	none	Supported

Outline



- Design Objectives
- SPARC T5 Processor Overview
- Core S3
- Cache Hierarchy Components
- Internode Coherency for 8-Socket Scaling
- Power Management Advances
- PCI-Express Gen3 I/O Subsystem
- Summary

Core Caches

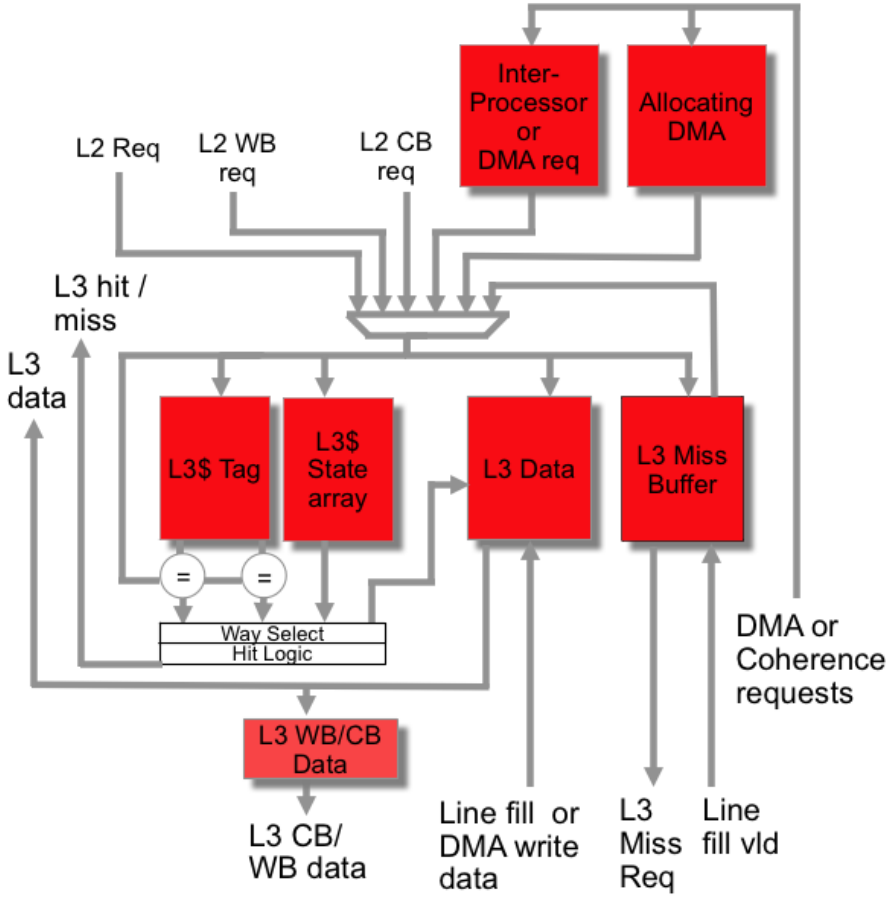
- 16 KB 4-way set associative L1 Instruction cache
- 16 KB 4-way set associative write through L1 Data cache
- 128 KB 8-way set associative, unified, inclusive L2 cache

L2-L3 Interconnect

- 8x9 Crossbar Switch connects the 16 cores to
 - 8 address interleaved address banks and
 - an I/O bridge
- The L3-L2 direction contains a control and data network.
 - control network provides a heads up for dependent instruction wake-up
 - Data network is used to return line fill data and send L3-L2 snoops
- Crossbar network has a bisection BW of 1 TBps, 2x T4

L3 Cache Overview

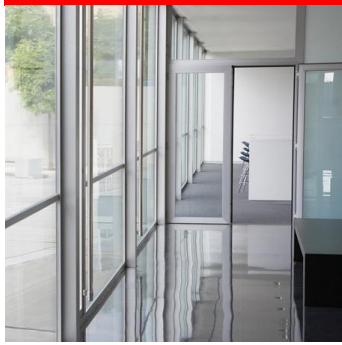
- 8MB (1MB per bank) 16-way set associative inclusive cache
- MOESI Coherence States per 64B cache line
- Tag results are used to
 - qualify data array access
 - align dependent issue with data return or conditionally flush thread
- Precise tag ”reverse” directory keeps L2s coherent



L3 Cache Overview (continued)

- Speeds up IO by allocating DMA buffers in the cache
 - Enhances clustered application performance
- Acceleration of contended locks
 - L3 forms a chain of same address requests
 - Processes them atomically on receiving an exclusive copy
- Supports coherent flushing and retirement of cache lines to avoid persistent errors

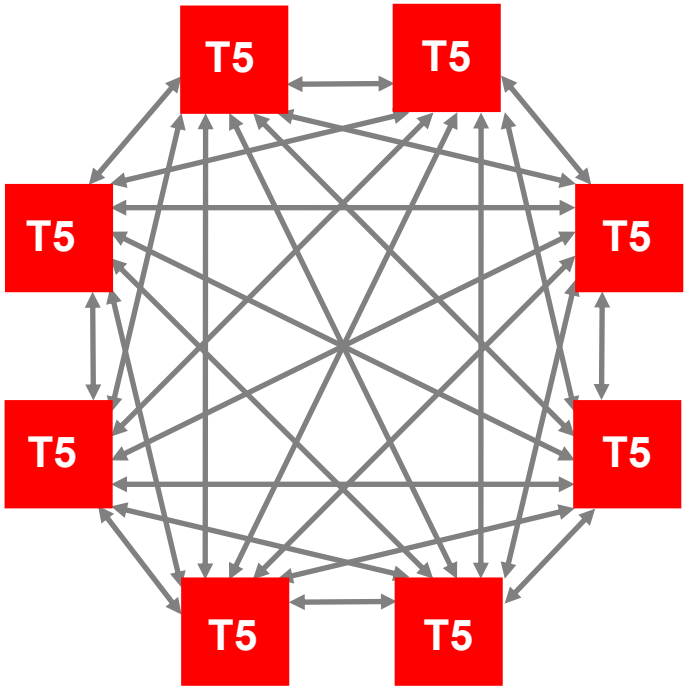
Outline



- Design Objectives
- SPARC T5 Processor Overview
- Core S3
- Cache Hierarchy Components
- Internode Coherency for 8-Socket Scaling
- Power Management Advances
- PCI-Express Gen3 I/O Subsystem
- Summary

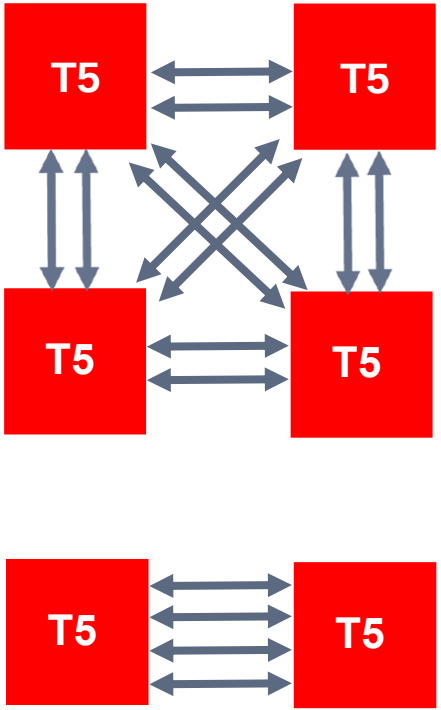
Internode Coherency Overview

- Glueless 1-hop scaling to eight sockets
- A precise directory tracks all L3s in the system
 - striped across all processors
 - stored in on-chip SRAMs
 - flexible for different socket counts
- Higher BW efficiency than snoop-based protocols enables better scaling
 - 50% more effective bandwidth than comparable snoopy implementation



Internode Coherency Fabric

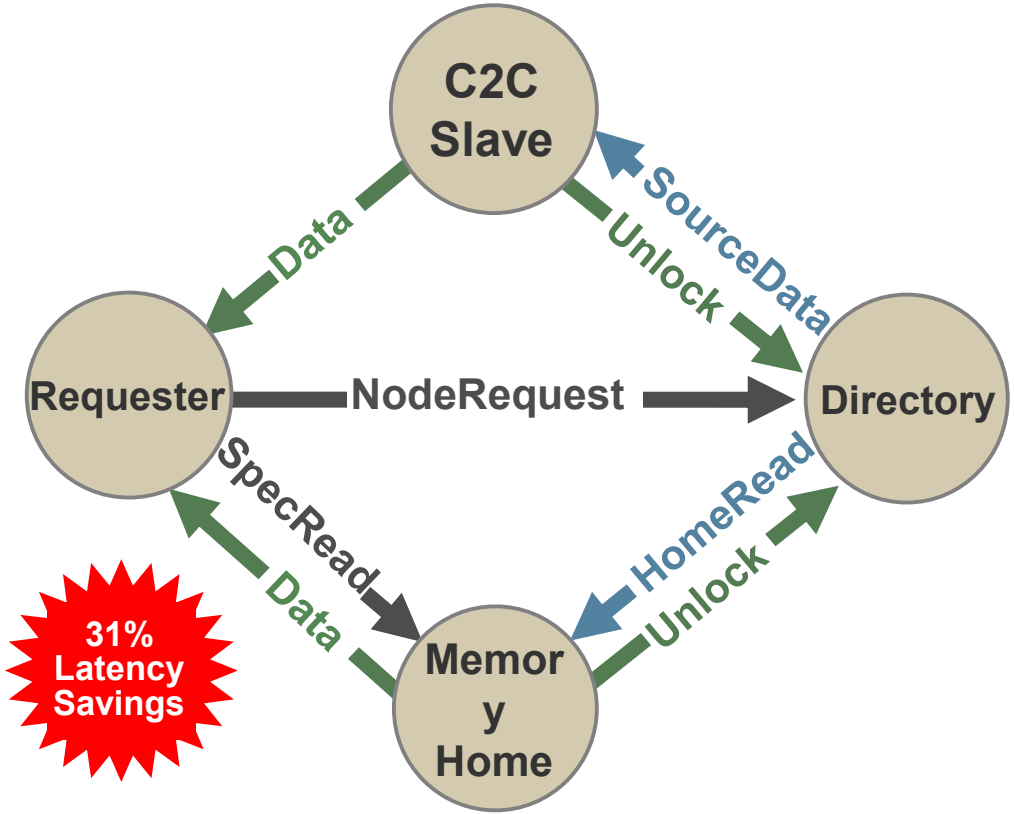
- Each link is 14 lanes wide and runs up to 15Gbps per lane
- Directly connected links minimize latency
- Trunked links achieve more bandwidth in smaller configurations
- Supports single lane failover



Internode Performance Optimizations

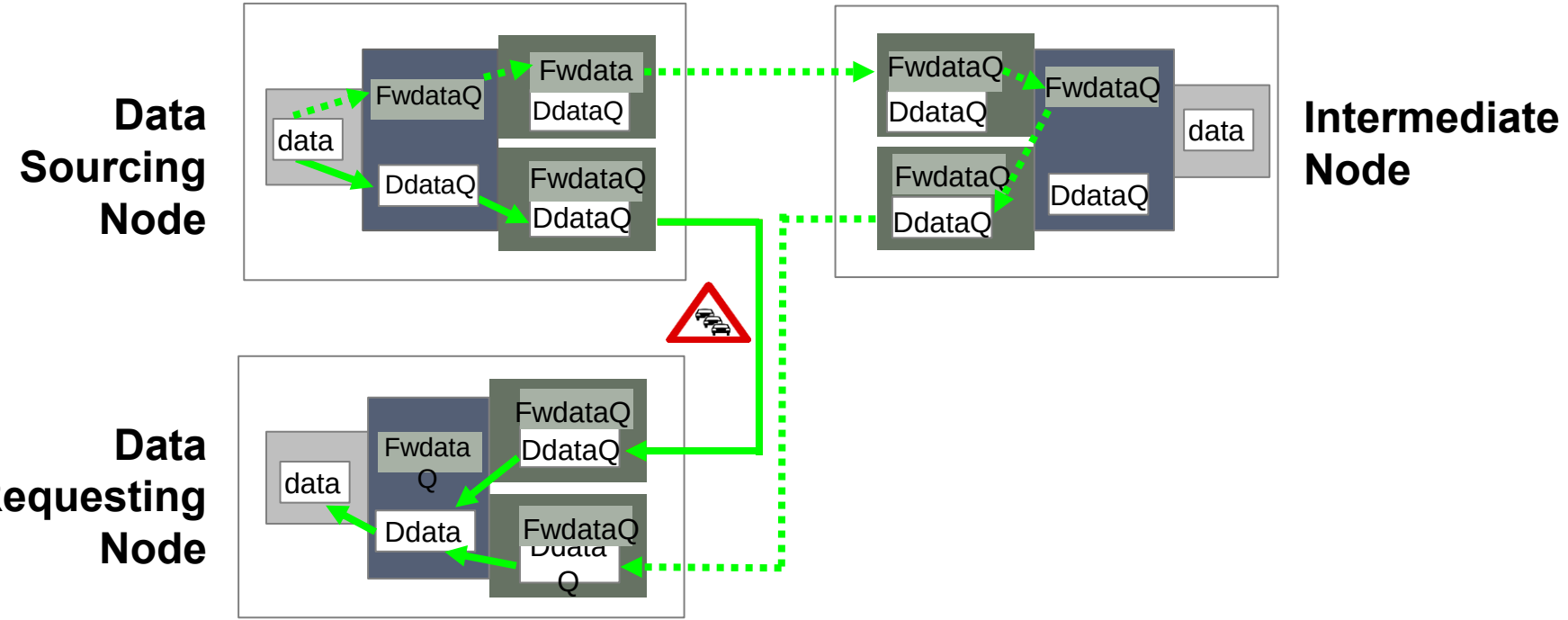
- Speculative memory reads prior to cache line serialization in the directory
- Cache-to-cache line transfers between nodes
- Dynamic congestion avoidance routes inter-node data around congested links

Internode Transaction Flow



1. A Requester issues a **NodeRequest** to the Directory and a **SpeculativeRead** to the Memory Home
2. After a Directory lookup, either a **HomeRead** or a **SourceData** request is generated
3. **Data** is returned from the Memory Home or C2C Slave

Dynamic Congestion Avoidance

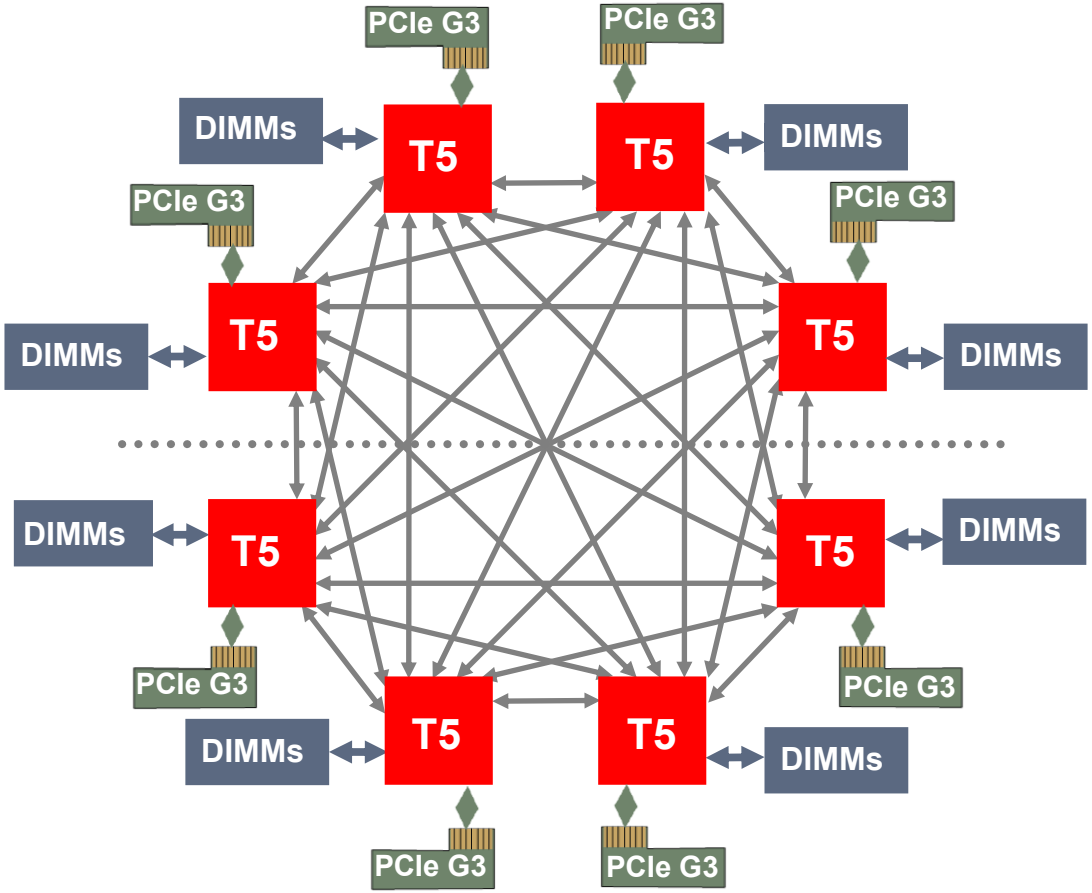


T5-8 Bandwidth

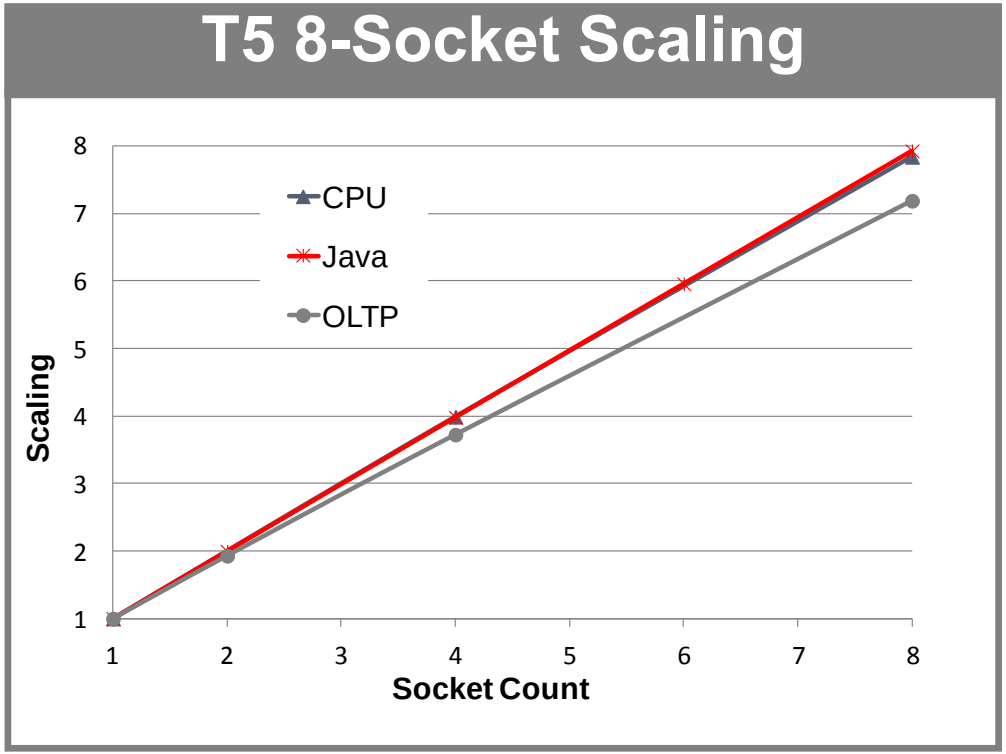
DDR3-1066 1+TB/sec

Coherency Bisection Bandwidth 840 GB/sec

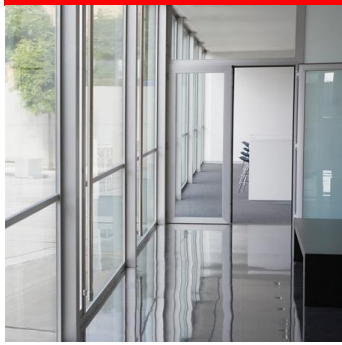
PCI Gen3 Bandwidth 256 GB/sec



Multiprocessor Performance



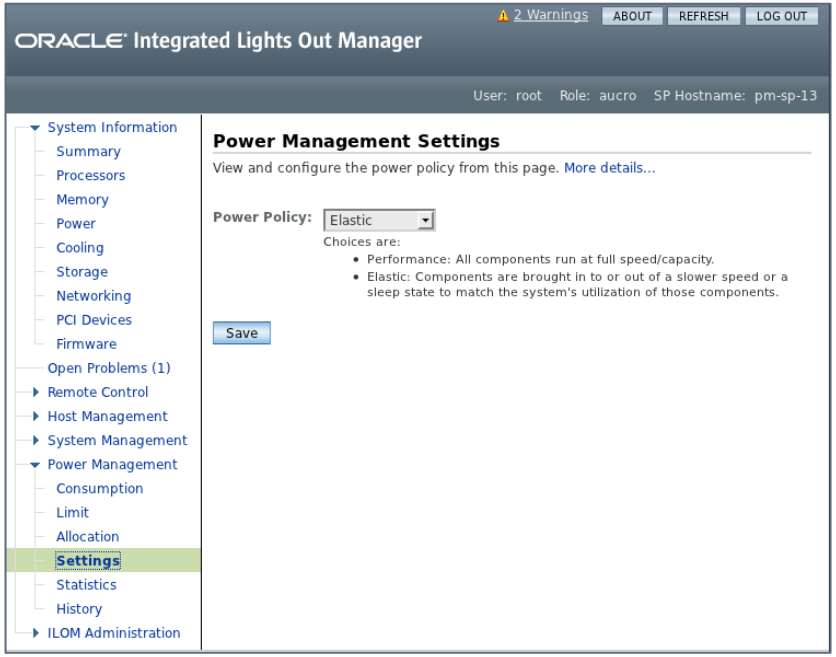
Outline



- Design Objectives
- SPARC T5 Processor Overview
- Core S3
- Cache Hierarchy Components
- Internode Coherency for 8-Socket Scaling
- Power Management Advances
- PCI-Express Gen3 I/O Subsystem
- Summary

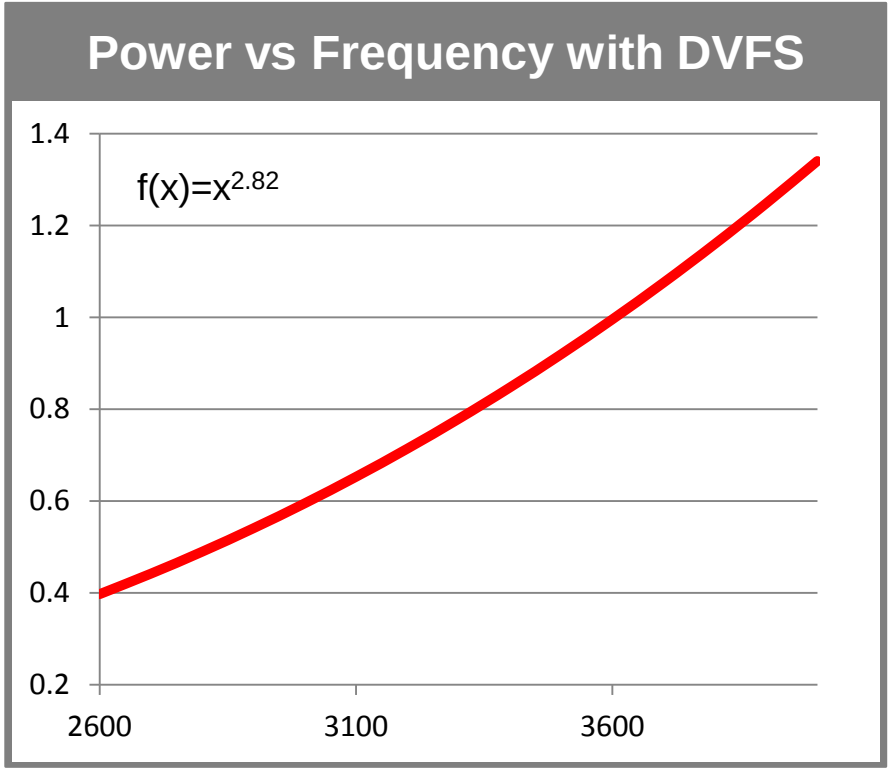
Power Management Advances

- Hardware saves power below 100% utilization with:
 - Chip wide DVFS
 - Per core pair cycle skipping
 - SerDes power scaling
 - DIMM off-lining w/ Dynamic Reconfiguration
 - DRAM PPSE and PPFE support
 - PCI Express Power Management
 - Clock Gating
- When peak performance is demanded
 - Power Management Controller achieves maximum frequency within customer imposed power and thermal limits



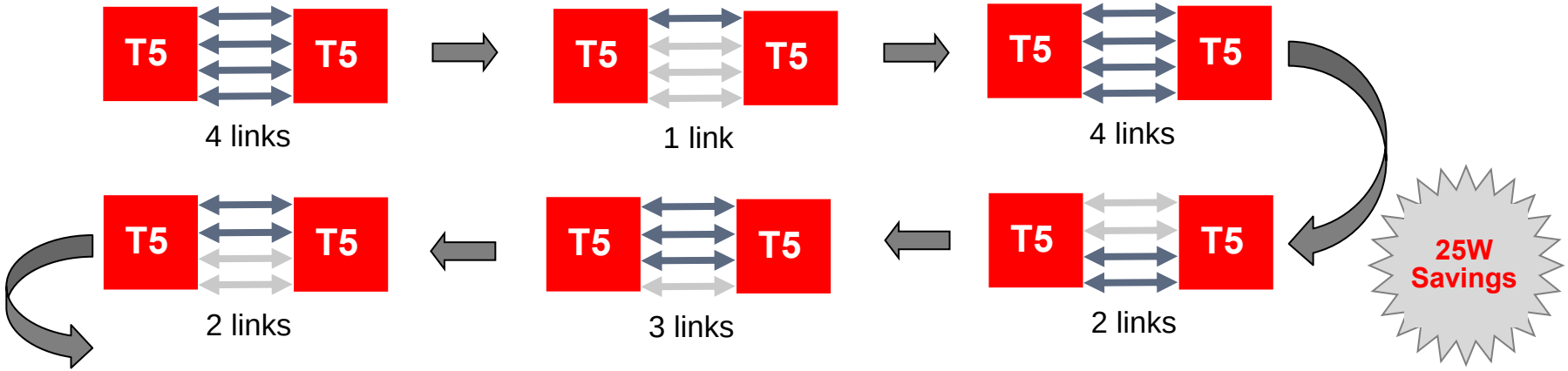
Power Management Controller: Elastic Savings

- Hardware saves power below 100% utilization
 - Chip wide DVFS
 - Per core pair cycle skipping
- Software monitors frequency needs of all cores
 - Puts chip at DVFS point satisfying all cores requirements
 - Puts core pairs at lowest cycle skip ratio satisfying 2 cores in the pair



Coherency Link Power Savings

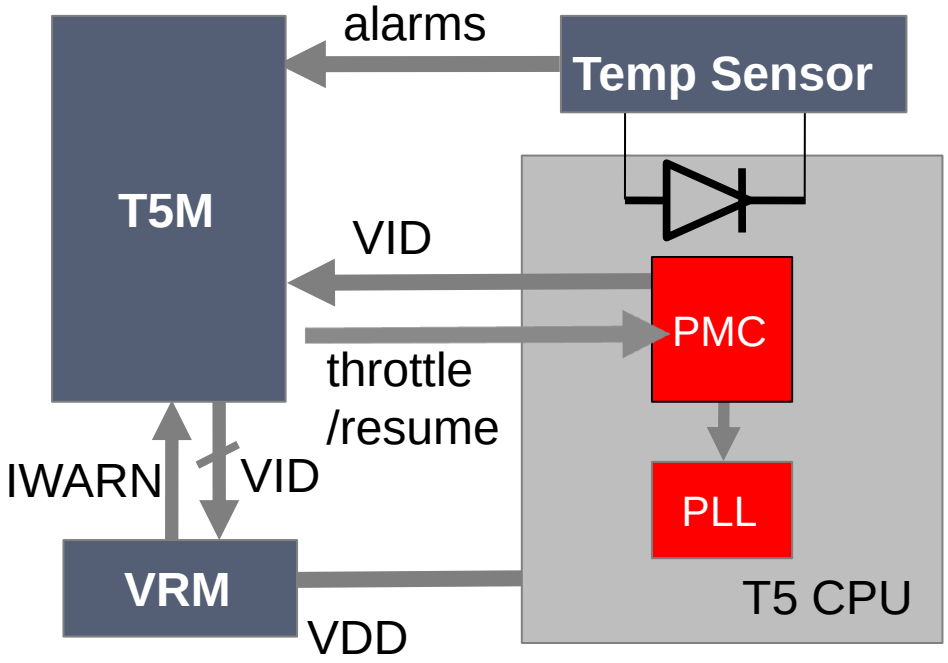
- Link scaling (4,3,2,1 dynamically as needed)
 - Hardware monitors link utilization
 - Software sets entry exit policy (thresholds and dwell times)



Memory Link Power Savings

- Two-levels of memory link standby
 - L0s: Power savings with fast wake up
 - Light sleep for N frames, then wake up and listen for data
 - L1: Much more power savings with longer wake up
 - Completely power off both tx and rx except for PLL
 - Used for unallocated memory regions

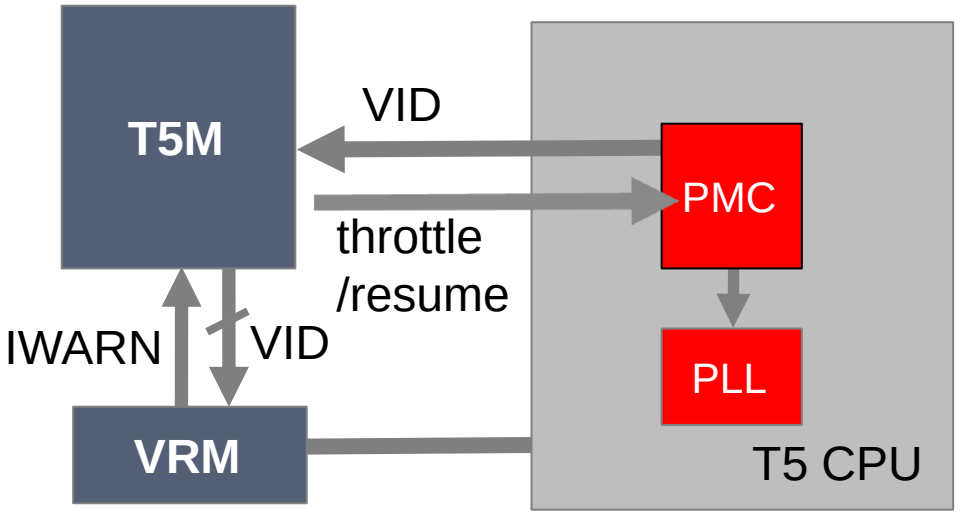
Peak Performance Thermal Management



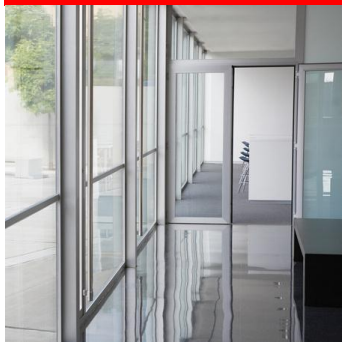
- 4 thermal diodes per chip
 - centered in core quads
- If any T > high-water mark
 Drop Freq, V
- If all T < low-water mark
 Raise Freq, V

Peak Performance Current Management

- Drop F,V if any current > high-water mark
- Raise F,V if any current < low-water mark
- Controls currents for CPU VDD plus motherboard and DIMMs



Outline



- Design Objectives
- SPARC T5 Processor Overview
- Core S3
- Cache Hierarchy Components
- Internode Coherency for 8-Socket Scaling
- Power Management Advances
- PCI-Express Gen3 I/O Subsystem
- Summary

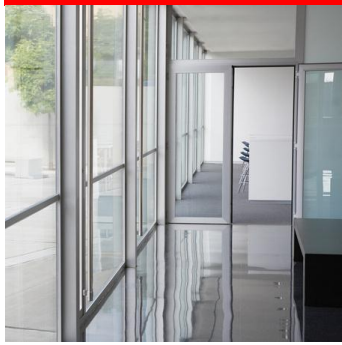
T5 PCIe Subsystem

- Dual x8 PCI Express Gen 3 ports provide 32 GB/s peak b/w
- Supports Atomic Fetch-and-Add, Unconditional-Swap and Compare-and-Swap operations
- Accelerates virtualized I/O with Oracle Solaris VMs
 - 128k virtual function address spaces ensure direct SR-IOV access for all logical domains
 - 64-bit DVMA space reduces IO mapping overhead, improving network performance
 - Guarantees fault and performance isolation among guest OS instances
- Supports PCI Express Power Management

T5 PCIe Progression

	T4	T5
PCI Express revision	Gen 2 (dual x8 ports)	Gen 3 (dual x8 ports)
Throughput full duplex	16 GBs	32 GBs
Data Management Unit	Single shared unit for both x8 PCIe ports	Two independent units one for each x8 PCIe port
Physical Address Support	44 bit	48 bit
Transaction Id Identification on MSI and MSI-X	No	Yes
PCIe Atomic Transactions	No	Yes
TLP Processing Hints	No	Yes, directs data to L3 cache
PCIe 2.0 compliance (ECN “Internal Error Reporting”)	Signaled via MSI interrupt	Signaled via PCIe message

Outline



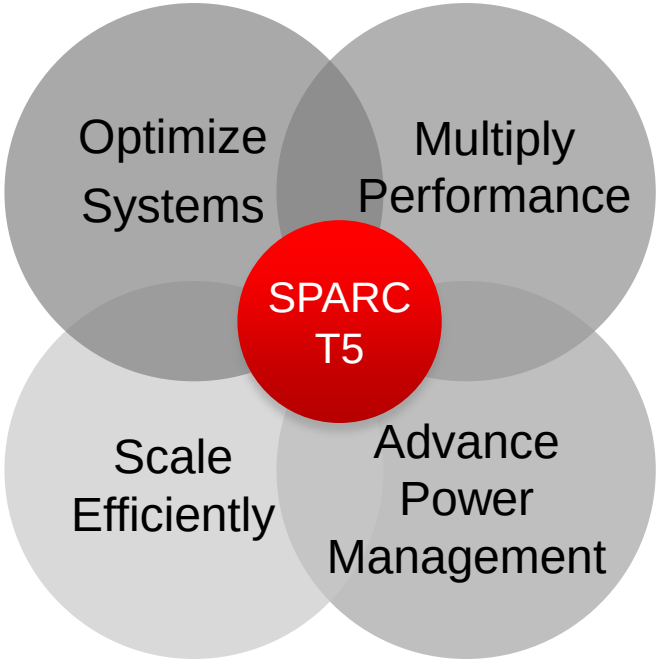
- Design Objectives
- SPARC T5 Processor Overview
- Core S3
- Cache Hierarchy Components
- Internode Coherency for 8-Socket Scaling
- Power Management Advances
- PCI-Express Gen3 I/O Subsystem
- Summary

SPARC T5 Summary

- Processor provides
 - Leadership throughput and per-thread performance
 - The industry's best on-chip encryption acceleration
 - Advanced power management
 - Highly-efficient one hop glueless scalability to 8 sockets
 - Enterprise-class general purpose computing and RAS
- SPARC T5 is the world's best processor for running Oracle software
 - Oracle Database, Fusion Applications, Fusion Middleware

Design Objectives Achieved

- Oracle workloads
- Engineered Systems
- Extends
 - ✓on-chip crypto acceleration
 - ✓RAS



- Double cores and cache
- Balance single thread and throughput
- Dynamically thread

- Scales to 8 sockets using directory
- Minimizes latency
- Avoids congestion
- Maximize bandwidth

- Maximizes peak performance
- Manages thermal and current loads
- Scales elastically

Q&A

Hardware and Software

ORACLE

Engineered to Work Together


ORACLE®

IBM

Smarter Systems for a
Smarter Planet

IBM zNext –

The 3rd Generation High Frequency Microprocessor Chip

Chung-Lung (Kevin) Shum
Senior Technical Staff Member, System z Processor Development, Systems & Technology Group, IBM Corp.



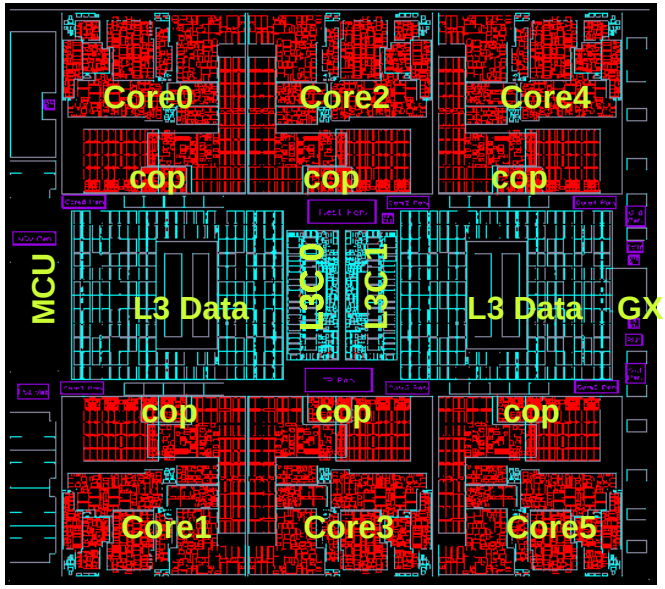
© 2012 IBM Corporation

zNext PU Chip Overview

- IBM mainframe microprocessor chip for the next generation of System z servers

- 32nm SOI technology
 - 597 mm2 (23.7mm x 25.2mm)
 - 15 layers of metal, 7.68 miles of wire
 - 2.75 Billion transistors
 - I/Os: 10000+ Power, 1071 Signal
 - SMP connections to external Hub chip (SC)
 - I/O Bus Controller (GX)
 - Memory Controller (MC) with **prefetching**

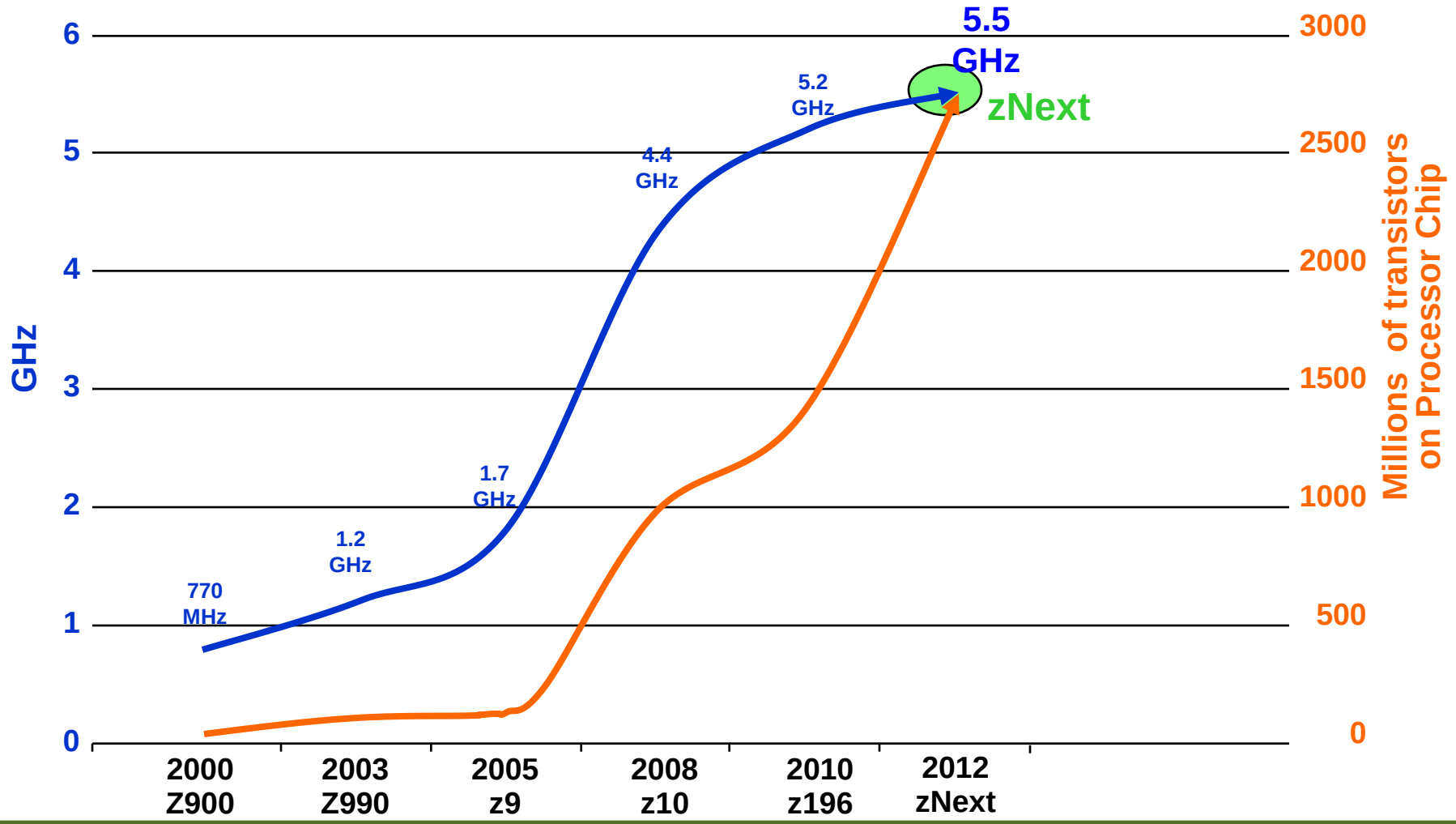
- Chip Features (vs. z196)
 - **6 new cores** per chip (vs. 4)
 - Core-**Dedicated** (vs. shared) Co-Processors
 - **48 MB EDRAM** on-chip shared L3 (2x)



- Processor Core Features
 - 2nd Generation out-of-order design
 - Speed & feed improvements
 - Microarchitecture innovations
 - Architecture extensions for software exploitations, e.g.,
 - Hardware Transactional Memory
 - Runtime instrumentation

© 2012 IBM Corporation

Speed: Higher Operating Frequency



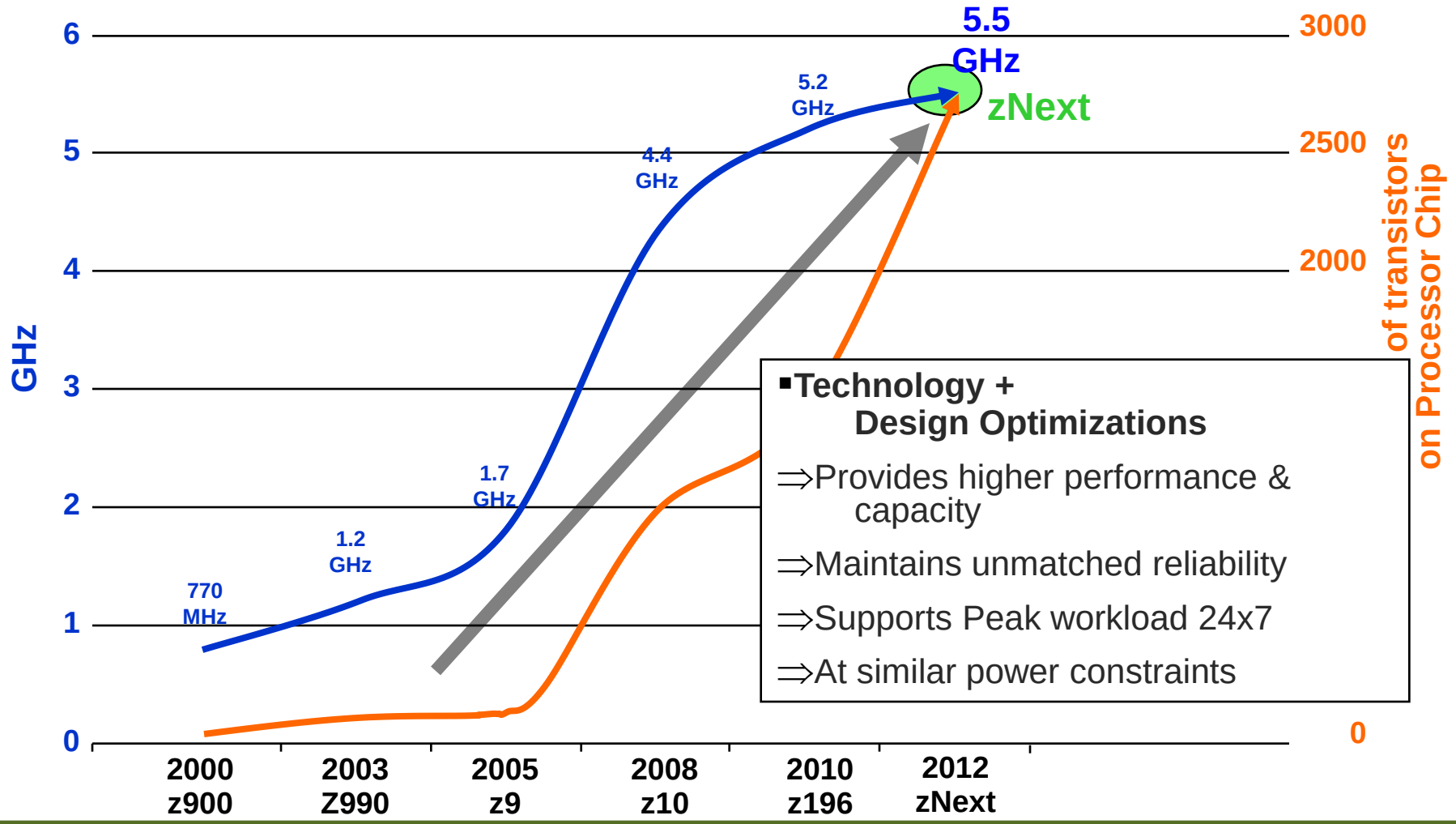
- **z900** – Full 64-bit z/Architecture
- **z990** – Superscalar CISC pipeline
- **z9** – System level scaling

- **z10** – Deep Pipeline, Arch. extensions
- **z196** – Out-Of-Order (OOO), Additional Architectural Extensions

- **zNext** – OOO+, Architectural Extensions, Enablement for new Software Paradigms

© 2012 IBM Corporation

Speed: Higher Operating Frequency



- **z900** – Full 64-bit z/Architecture
- **z990** – Superscalar CISC pipeline
- **z9** – System level scaling

- **z10** – Deep Pipeline, Arch. extensions
- **z196** – Out-Of-Order (OOO), Additional Architectural Extensions

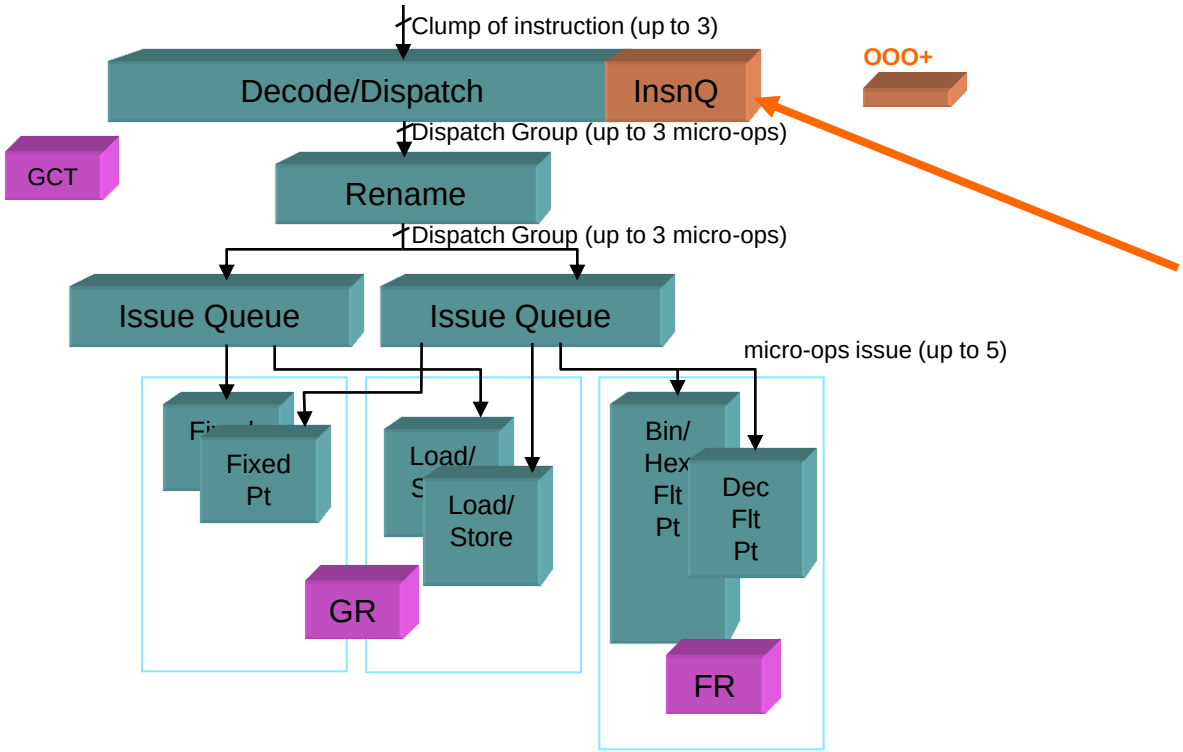
- **zNext** – OOO+, Architectural Extensions, Enablement for new Software Paradigms

© 2012 IBM Corporation

Feed Improvements: Maximizing Out-of-Order Window

- Improved dispatch grouping efficiencies
 - ➔ More instructions per group
 - – Reduced cracked instructions overhead
 - – Increased branches per group to 2
 - – Added Instruction Queue (InsnQ) for re-clumping

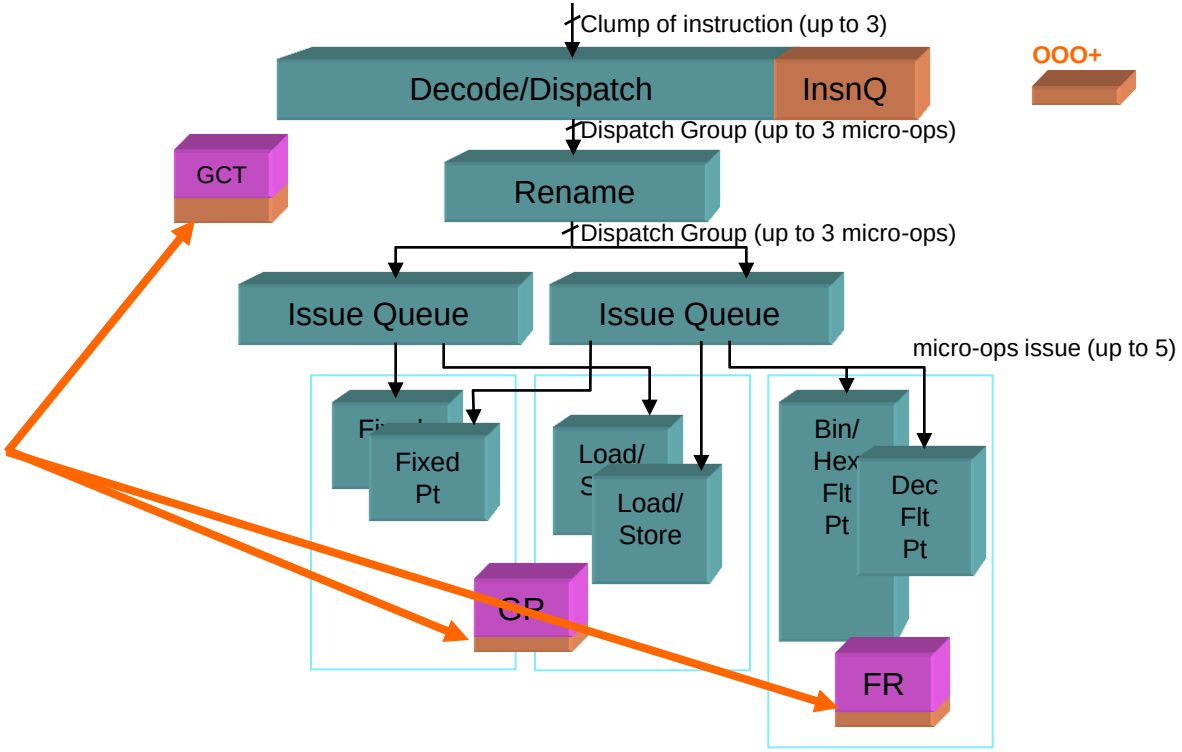
*clumps – parcel of instructions delivered from instruction fetching



© 2012 IBM Corporation

Feed Improvements: Maximizing Out-of-Order Window

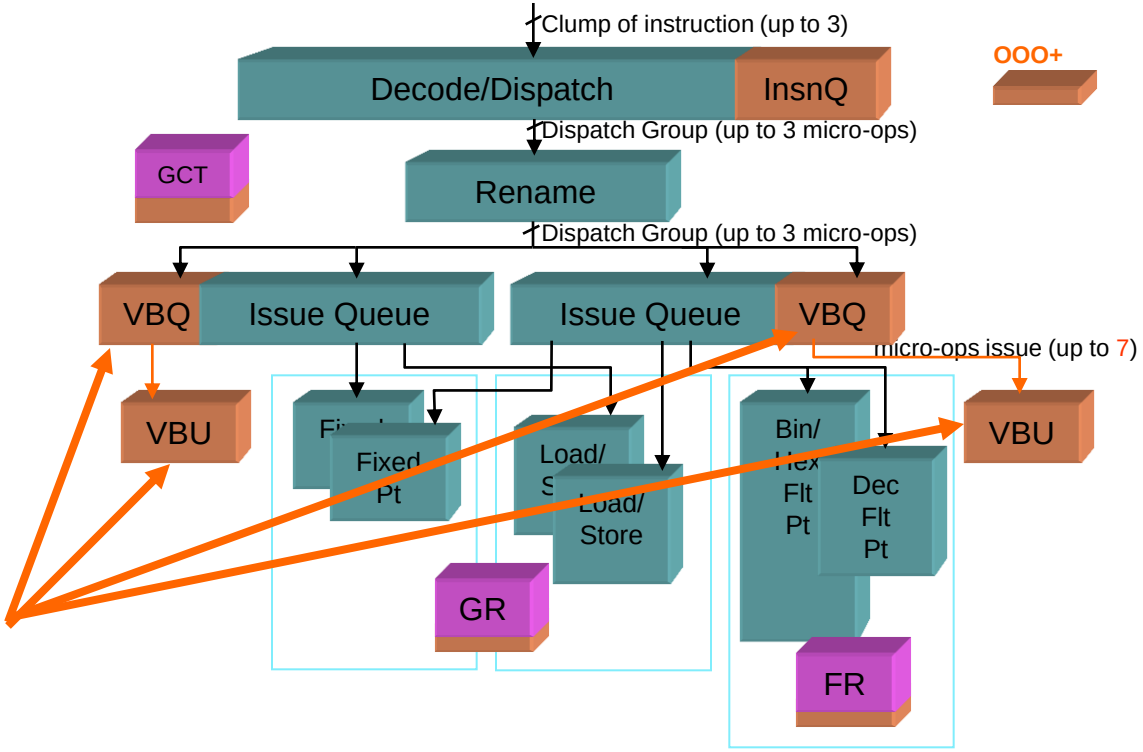
- Increased out-of-order resources
 - ➔ More out-of-order groups
 - – Multi-grouped instructions speculative completion
 - – Increased Global Completion Table (GCT) entries to 30x3 (+25%)
 - – Increased usable physical GR entries to 80 (+25%)
 - – Increased physical FR entries to 64 (+33%)



© 2012 IBM Corporation

Feed Improvements: Maximizing Out-of-Order Window

- Increased execution bandwidth
 - ➔ More instructions issued per cycle
 - – Added Virtual Branch Queue (VBQ) for relative branch queuing
 - – Added Virtual Branch Unit (VBU) for relative branch execution
 - – Increased effective issue queue size to 32x2 (+60%)
 - – Increased issue bandwidth per cycle to 7 (+40%)



© 2012 IBM Corporation

z196 Pipeline (recap)

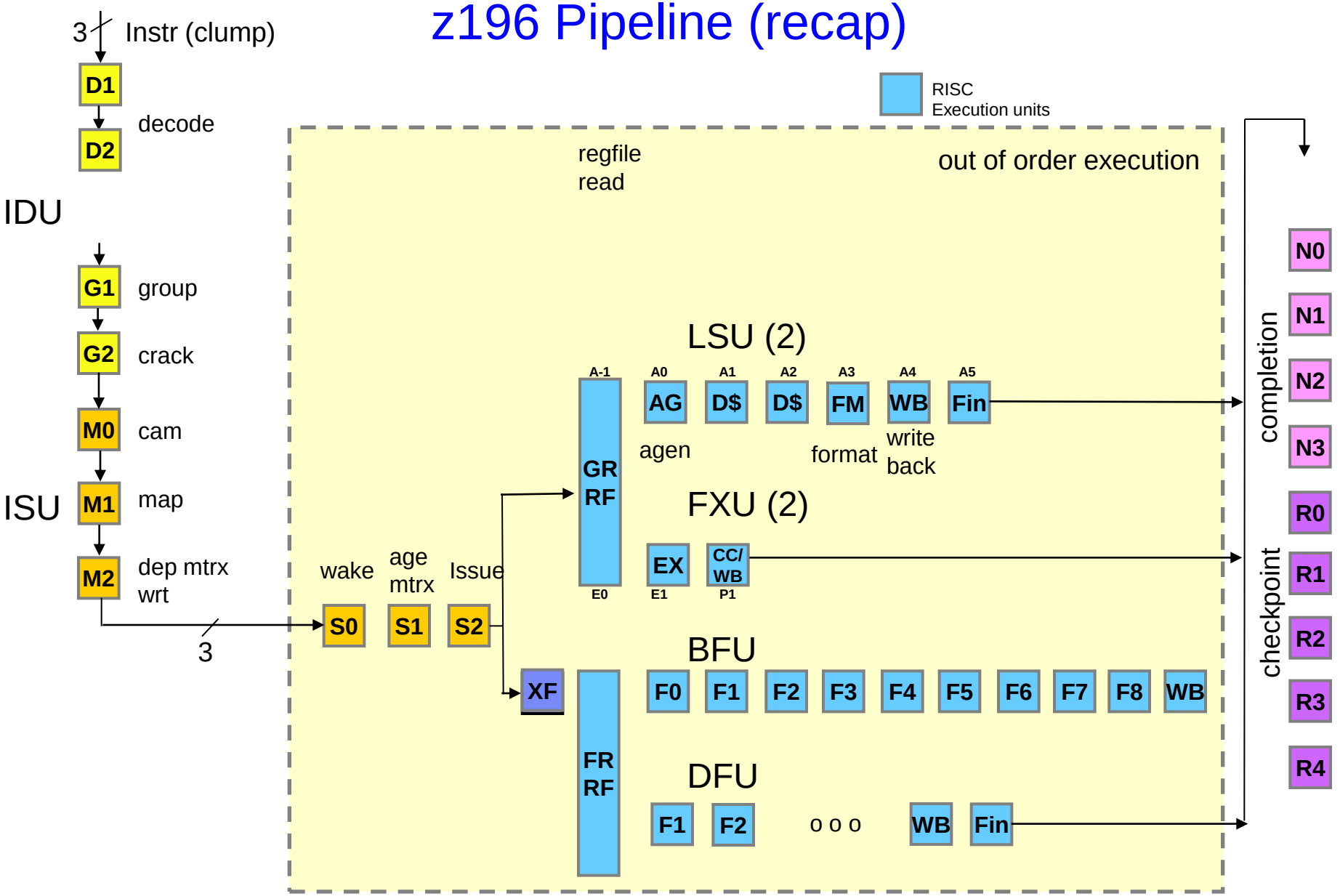
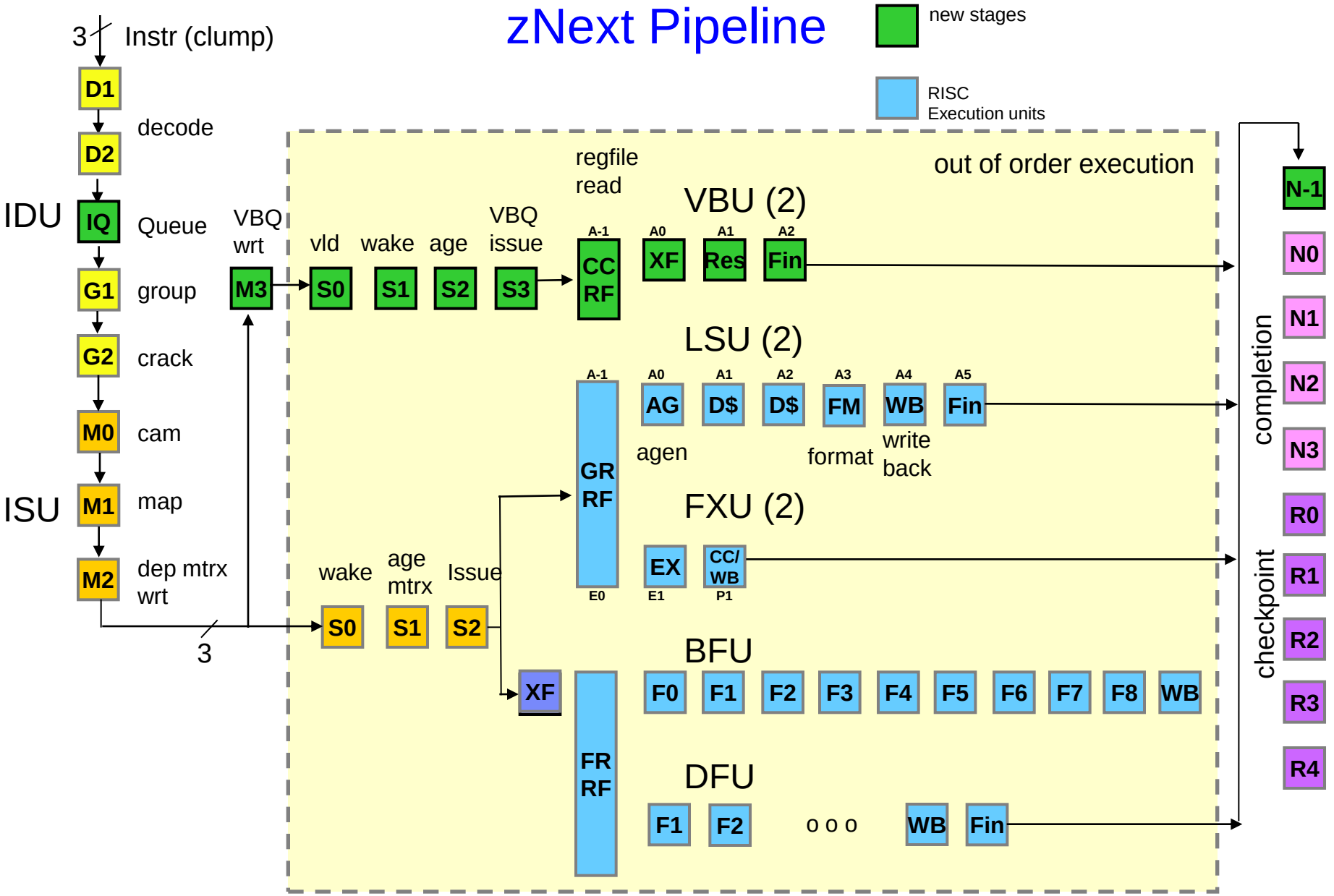


Diagram based on Brian Curran's HOTCHIP 22 presentation

© 2012 IBM Corporation

zNext Pipeline



© 2012 IBM Corporation

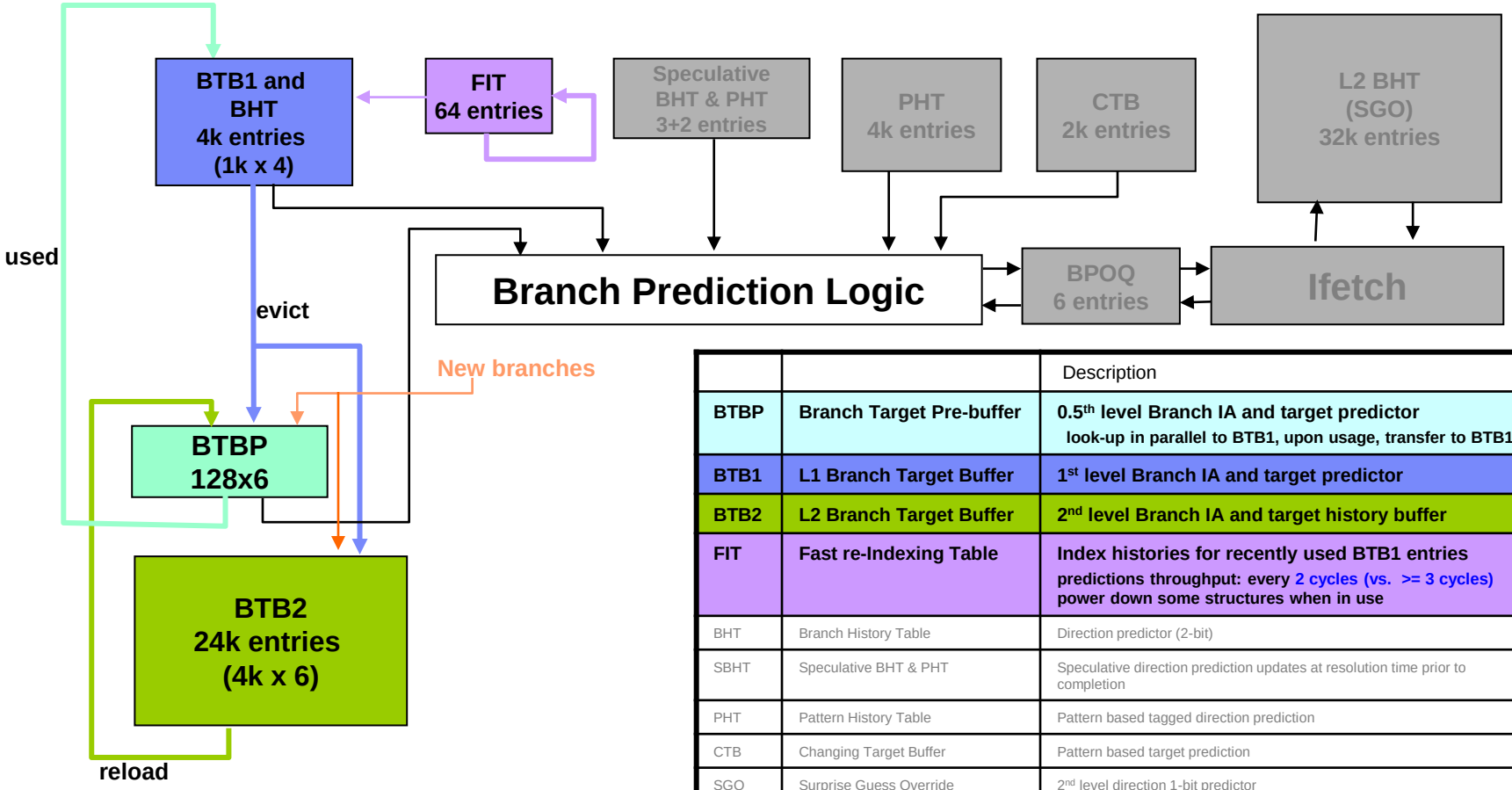
Feed Improvements: Accelerating Specific Functions

- Short-circuit executions
 - Common idioms executed during dispatch
 - e.g. initializing a GR with zeros
- D-Cache (SRAM design) with banking support
 - 32 banks for concurrent 2 read and 1 write operations
 - Faster cache writes reduce future load-use delays
- Dedicated Fixed-point divide engine resulting in 25-65% faster operations
- Millicode (Vertical Microcode) operations
 - Selective hardware execution
 Translate, Translate and Test, Store Clock
 - Shorter startup latency
 Move Character variations, Co-Processor operations
 - Hardware assists for prefetching (target cache level & coherency state)
 Move Character Long variations
 - Dedicated hardware for Unicode conversion (UTF8 <> UTF16)

© 2012 IBM Corporation

Micro-Architecture Innovations: Branch Prediction

- Branch Prediction is essential in improving performance
 - 2nd level BTB (BTB2) for capacity (more than 3x)
 - Fast re-Indexing Table (FIT) for latency (up to 33% reduction)



		Description
BTBP	**Branch Target Pre-buffer**	**0.5th level Branch IA and target predictor** look-up in parallel to BTB1, upon usage, transfer to BTB1
BTB1	**L1 Branch Target Buffer**	**1st level Branch IA and target predictor**
BTB2	**L2 Branch Target Buffer**	**2nd level Branch IA and target history buffer**
FIT	**Fast re-Indexing Table**	**Index histories for recently used BTB1 entries** predictions throughput: every 2 cycles (vs. >= 3 cycles) power down some structures when in use
BHT	Branch History Table	Direction predictor (2-bit)
SBHT	Speculative BHT & PHT	Speculative direction prediction updates at resolution time prior to completion
PHT	Pattern History Table	Pattern based tagged direction prediction
CTB	Changing Target Buffer	Pattern based target prediction
SGO	Surprise Guess Override	2nd level direction 1-bit predictor

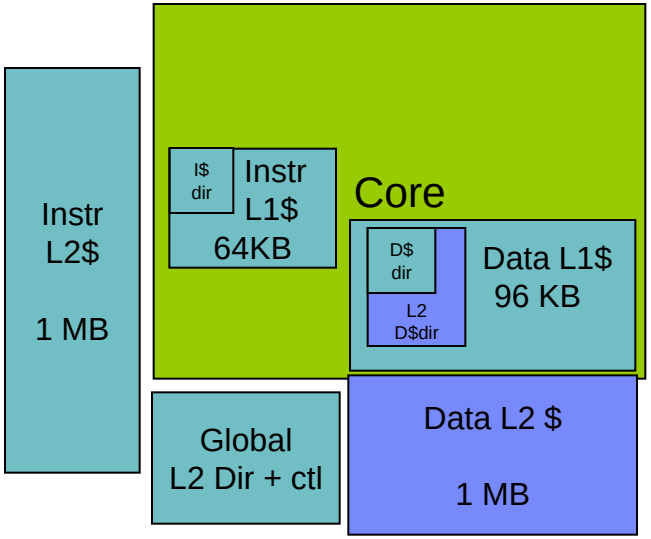
Diagram & Table based on Eric Schwarz's 2011 VAIL presentation

© 2012 IBM Corporation

Micro-Architecture Innovations: Cache Subsystem

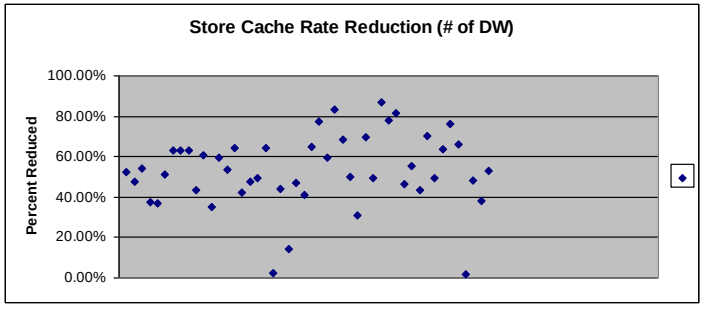
- Split Level 2 Cache (instead of unified)
 - 1M-byte Instruction, 1M-byte Data
 - Inclusive of instruction-L1 (64 Kbyte) and data-L1 (96 Kbyte)
 - Bigger aggregate L2 with shorter latency
- Integrated data-L2 directory
 - data-L2 directory is merged into data-L1 directory
 - Logically indexed like in data-L1 directory
 - L2 Hit / miss knowledge at L1 miss time

➔ L1 miss, L2 hit latency reduced by up to 45%



- Store "Gathering" Cache
 - circular queue of 64 entries of half-lines (128 bytes)
 - merges stores to same half-line post L1 updates
 - reduces pipeline usage for stores in L2 and L3
 - Hardware Transactions storage updates

➔ Store traffic to L3 typically reduced by ~50%



Modeling Data provided by Jim Mitchell @ IBM Poughkeepsie

© 2012 IBM Corporation

Targeted Architectural Extensions

- 2 Gigabyte Page support

- Decimal Floating Point Extension
 - Instructions to convert numeric data between 2 formats:
 zoned fixed-point decimal, and
 decimal floating point

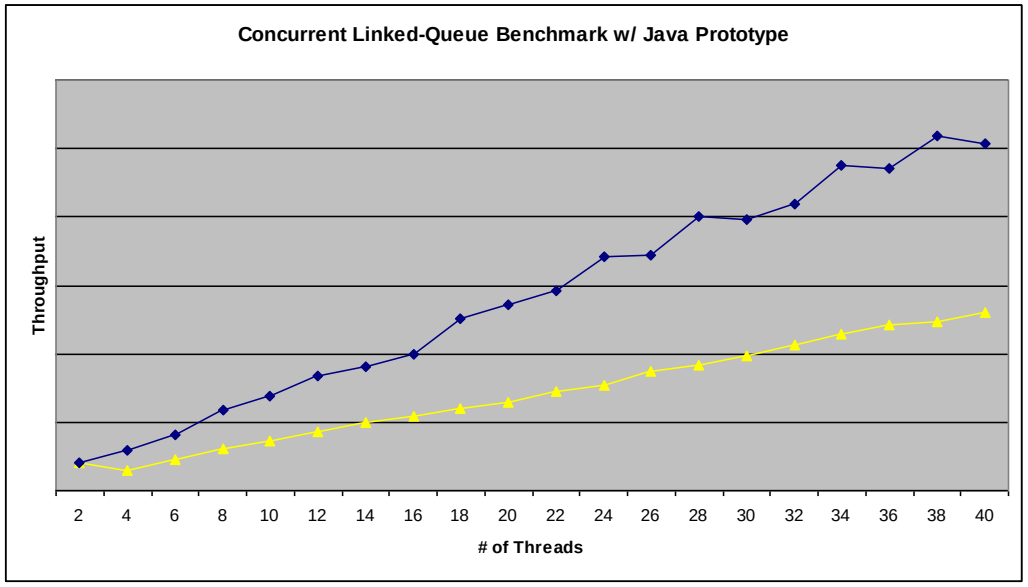
- Instruction Processing Directives
 - Branch preload instructions
 Specifies the address of a branch instruction and its target to be installed into branch prediction tables (through BTBP)

 - Data access intent instruction
 Specifies what operands of the next instruction may be further accessed for
 e.g. getting a cache line exclusive on a load for future store
 e.g. keeping access-once line at current Least-Recently-Used (LRU) position

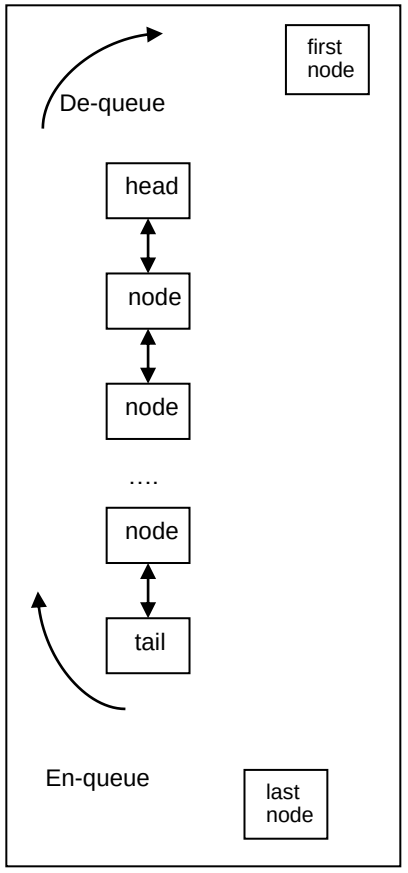
© 2012 IBM Corporation

Architectural Differentiation Extension: Transactional Execution

- General Purpose Multiprocessor Support
 - Instructions specifying start, end, and abort of a transaction
 - Pending storage updates are “shielded” from other processors until transaction completes
 - Implemented at heart of CPU (core+L1) for performance
 - Heavy focus on support for software usage and debug
 - “Constrained Transaction” with hardware auto-retries for code simplification
- Prototype benchmark with HTM
 - Showed ~2x improvements and better scalability (slope)



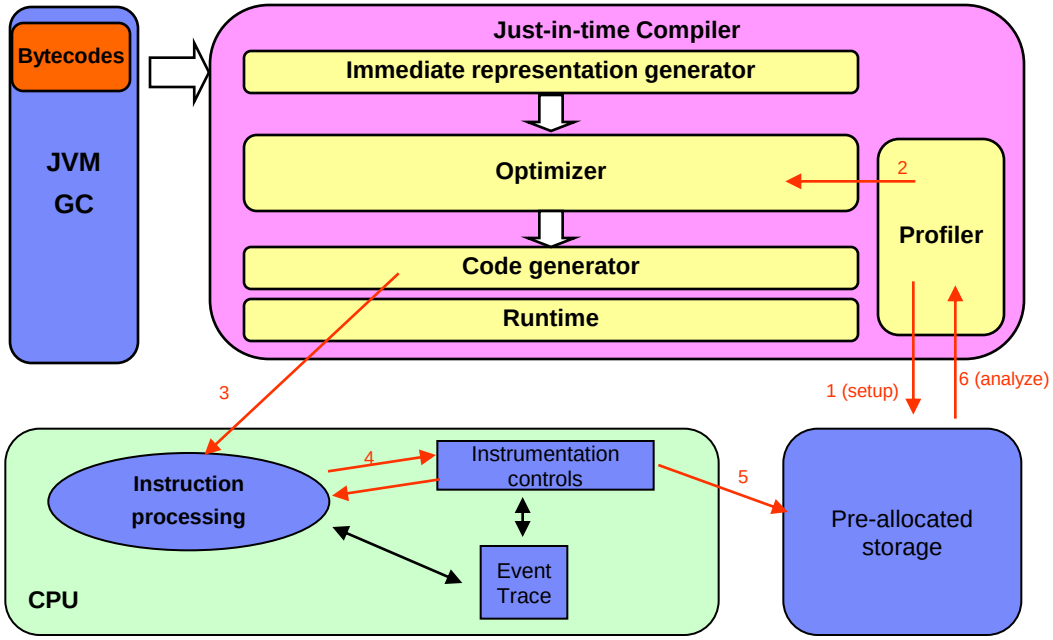
Prototype Data provided by Jerry Zheng, Marcel Mitran @ IBM Toronto

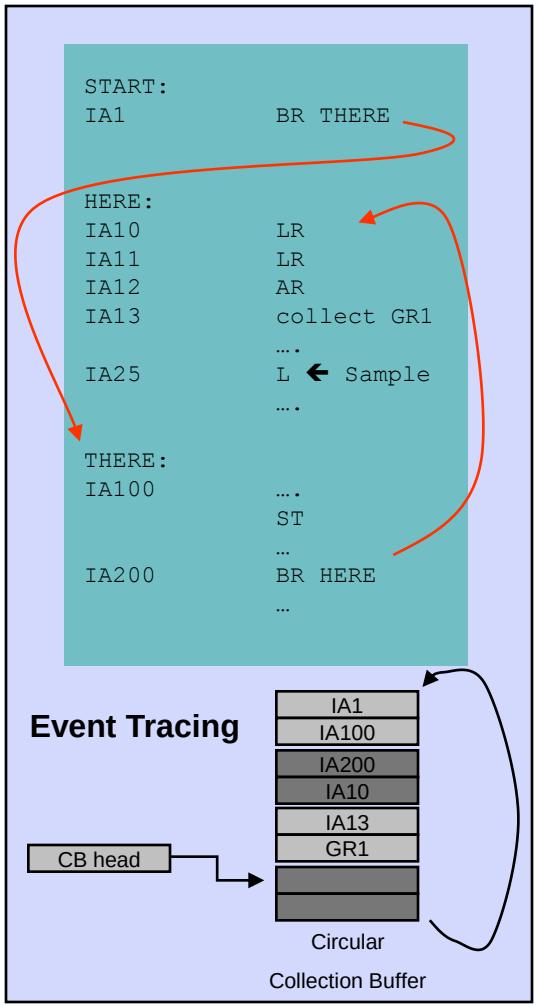


© 2012 IBM Corporation

Architectural Differentiation Extension: Runtime Instrumentation

- Low overhead profiling with hardware support
 - Instruction samples by time, count or explicit marking
- Sample reports include hard-to-get information:
 - Event traces, e.g. taken branch trace
 - "costly" events of interest, e.g. cache miss information
 - GR value profiling
- Enables better "self-tuning" opportunities

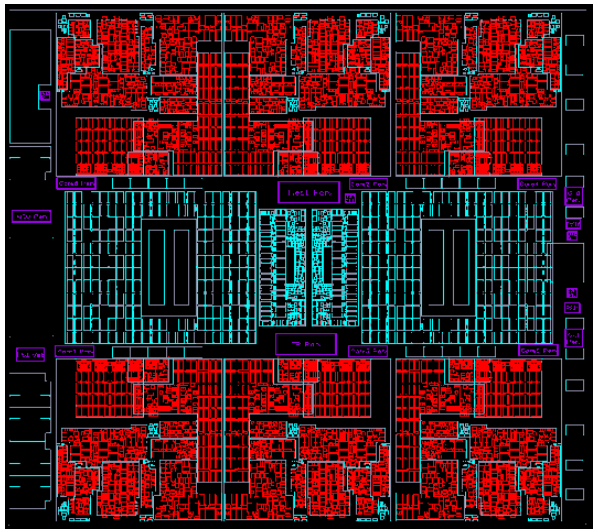




© 2012 IBM Corporation

Summary: zNext will…..

- Be used in a new family of IBM System z mainframe servers
- Sustain IBM's mainframe leadership in computing capacity and performance without sacrificing any reliability, with
 - Up to 6 active cores per chip
 - 48M-byte shared on-chip L3 cache
 - uniquely designed low-latency private L2 cache
 - >24K target and >32k direction branch histories
 - Numerous micro-architectural enhancements*
- Provide architecture extensions*, and be the 1st general purpose microprocessor to support
 - hardware transactional memory
 - software self-directed run-time profiling
- Be amongst the fastest microprocessors @ 5.5 GHz
 - joining z196 @ continuous clock-speed of >5 GHz



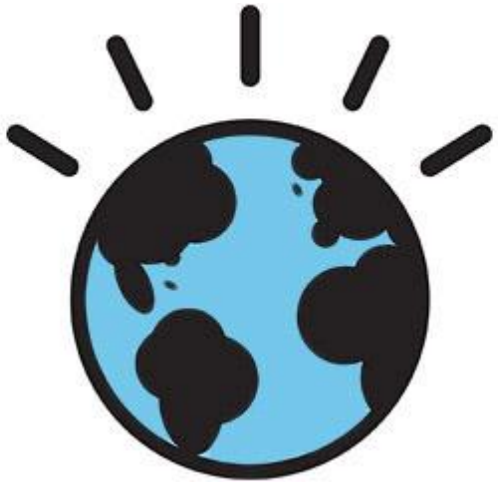
* Not all features and extensions described in this presentation

© 2012 IBM Corporation

Acknowledgements

- Microarchitecture, Design and Verification Team
 - Members from
 Austin, Bangalore, Boeblingen, Haifa, Poughkeepsie, Tel Aviv, and other design labs around the world
- Architecture, Software, Performance and Research Team
 - Members from
 z/OS, z/VM, z/Linux, Compiler, JAVA, DB2, etc.
- Project Management and Technical Executives

Thank You!



© 2012 IBM Corporation

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.

AIX*	FICON*	Parallel Sysplex*	System z10
BladeCenter*	GDPS*	POWER*	WebSphere*
CICS*	IMS	PR/SM	z/OS*
Cognos*	IBM*	System z*	z/VM*
DataPower*	IBM (logo)*	System z9*	z/VSE*
DB2*			zEnterprise*

* Registered trademarks of IBM Corporation

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license there from.
Java and all Java-based trademarks are trademarks of Oracle Corporation in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
InfiniBand is a trademark and service mark of the InfiniBand Trade Association.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
UNIX is a registered trademark of The Open Group in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency, which is now part of the Office of Government Commerce.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

© 2012 IBM Corporation

High Performance State Retention with Power Gating applied to CPU subsystems – design approaches and silicon evaluation

David Flynn, Fellow, R&D
ARM Ltd, Cambridge, UK
david.flynn@arm.com

ABSTRACT

Power management is of increasing concern and challenge to SOC and product designers [1], [2]. Power Gating (PG) is now well understood as a technique for reducing static leakage power when circuits are idle [3]. State-Retention Power Gating (SRPG) enhancements in hardware [4] can address fast wake-up latency and transparency to system software but have area, performance and robustness/reliability impacts that need minimizing [5].

This presentation addresses practical application of State Retention Power Gating to CPU subsystems, (but applicable to other SOC sub-systems) and covers what matters from the system and RTL designer perspective building on the EDA implementation support from UPF [6] and CPF [7] power intent.

Current EDA support for Power Gating is tuned around "logic-level" drive of power gates. The new techniques that are described and contrasted build on the multi-voltage aware tools and formats to add enhanced power gate performance as well as addressing state retention without the traditional area and timing penalties.

The work described in this paper is at an applied research phase and has been undertaken in collaboration with researches in the Electronics and Computer Science faculty of the University of Southampton in the UK; the technology demonstrator implemented in Silicon (on a 65nm Low Leakage process) was co-developed and fabricated using the EUROPRACTICE "mini@sic" Multi-Project Wafer service with TSMC Inc as the semiconductor foundry [8].

1. BACKGROUND

The research group at ARM has worked for a number of years with customers and leading EDA partners to take complex 'expert' low-power industry techniques and facilitate their successful adoption for standard System-on-Chip designers and implementers. This increasingly requires the development of Physical IP components and model abstractions that support the current and evolving Multi-Voltage tools and UPF and CPF standards.

2. BUILDING ON BASIC POWER GATING

The multi-voltage tools support and associated power intent now prove to be a foundation for more advanced techniques to improve on the base-line power-gating and state retention support envisaged as the EDA tools were developed [9].

2.1 Multi-Voltage Power Gating

Industry standard "Multi-Voltage" EDA tools support logic level drive of the gate terminal of the power switches, while more expert approaches have traditionally been required to add Gate Bias to improve the off-current (ratio) of power gates [10].

A Super-Cutoff CMOS "buffered" power gate cell family with integrated level shifting has been developed to work seamlessly with standard EDA MV tool flows (shown in figure 1). Header power gates are of primary interest to facilitate simple generation of gate bias supply voltage (the core voltage rail augmented by small charge pump or regulated from a higher IO supply rail).

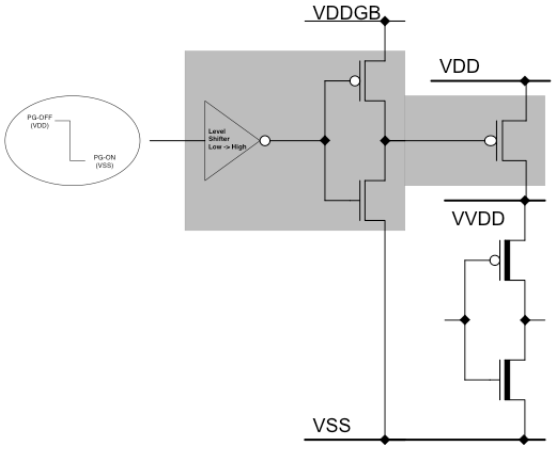


Figure 1: SCCMOS enhanced power gate

The multi-voltage internals of the enhanced switch are hidden from the implementation tools and support lower off-current with High-Vth "MTCMOS" power switches, or lower-IR drop with standard Vth switches.

3. ENHANCING BASIC STATE RETENTION

The experimental approach adopted has been to amortize the cost of state retention across multiple registers by splitting the power rails for high performance flip-flops (a near-zero area cost) and amortize the retention cost by managing the clamping of clocks and resets efficiently in the SOC implementation flow such that the speed and area impacts are minimized over and above the cost of Power Gating that designers well understand.

Figure 2 illustrates how the retention power domain is distributed to manage "live-slave" state retention between clock-gates and registers. For registers with asynchronous reset controls such controls must also be explicitly clamped similarly.

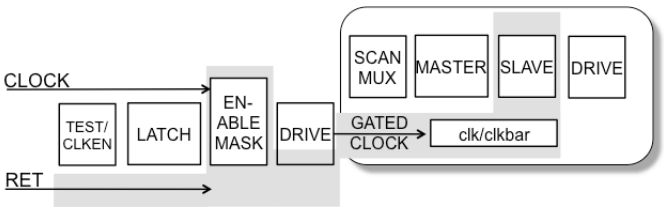


Figure 2: Advanced SRPG distributed retention domain

For short-term SRPG support the slave latches and associated clamping domains must be kept powered. For deep sleep this domain (shown with gray overlay) would be power-gated off as well (state lost PG, potentially requiring software).

Voltage scaling of the state retention rail is attractive to provided an extended SRPG mode of operation, but simple techniques such

as adding a Vt-drop that was safe at higher-voltage process nodes [4] do not provide sufficient safe state-integrity margin for latch structures on sub 90nm technologies with higher inter-device variation on latch feedback structures. Figure 3 shows the addition of a Boosted-Gate "drowsy" retention to the buffered SCCMOS power-gate of Figure 1 where the raised-voltage Gate Bias supply provides additional headroom to the scaled retention voltage.

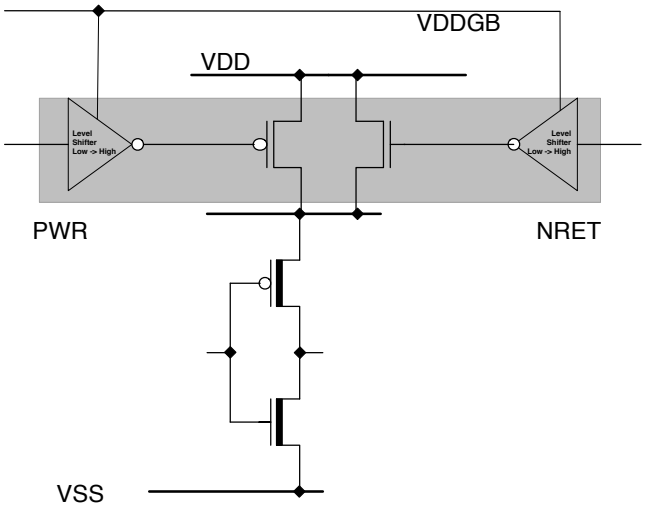


Figure 3: SCCMOS with Boosted-Gate retention

State retention needs to be 100% robust and reliable in the presence of power-gating transients and noise from neighboring blocks that share a common ground or supply. The underlying retention registers need to be designed to balance retention leakage power with safe retention latch structures. The poster describes the experimental structures designed and implemented to evaluate and characterize the integrity of retention registers at reduced voltages.

4. TECHNOLOGY EVALUATION

Figure 4 depicts the layout of the test silicon implemented to validate the physical IP cell abstractions and EDA flow compatibility.

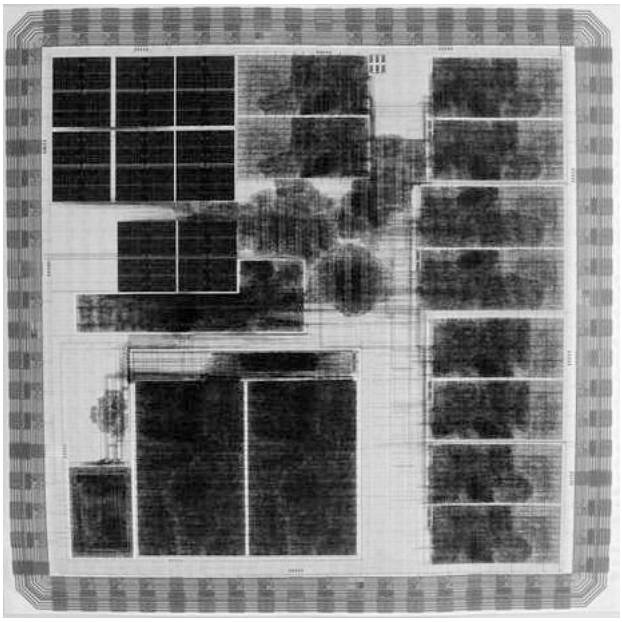


Figure 4: Advanced ARPG test silicon (TSMC65LP)

Due to small silicon die-size availability for rapid prototyping (2 x 2mm!) small ARM® Cortex-M0 ™ CPU macro-cells were chosen and constrained for performance to a worst case corner signoff at 330MHz, on the 65nm Low-Leakage technology. Five matched pairs of CPUs were instantiated with the 4 pairs on the right of the layout to evaluate standard PG and SRPG implementations plus the enhanced retention ARPG and SCCMOS plus DRPG gate bias implementations. The "tracking-pair" approach allows the implementations to be evaluated at 400MHz+ with each CPU of a pair having critical paths stressed in even and odd clock cycles, while the main SOC runs reliably zero-waits state at 200MHz. Finally, the chip includes state integrity structures in the lower-left layout to analyze state integrity and reliability in the presence of switching noise and power gating inrush.

REFERENCES

[1] Mudge, Trevor, "Power: A First Class Architectural Design Constraint" IEEE Computer, vol. 34, no. 4, April 2001. http://doi.ieeecomputersociety.org/10.1109/2.917539

[2] Keating, M., Flynn D. et al "Low Power Methodology Manual - for System-on-Chip Design", Springer 2007 ISBN: 978-0-387-71818-7 http://www.lpmm-book.org/

[3] Shi, K. Flynn, D. "Power Gating Design Tradeoffs and Considerations in Production Low-Power Designs", DesignCon 2009 http://www.designcon.com/infovault/paper.asp?PAPER_ID=474

[4] Mutoh S. et al. "A 1v multi-threshold voltage CMOS DSP with an efficient power management technique for mobile phone applications" ISSCC1996, pages 168–169, 1996.

[5] Flynn, D.. Gibbons, A. "Design for State Retention: Strategies and Case Studies" SNUG San Jose 2008, Track TA2

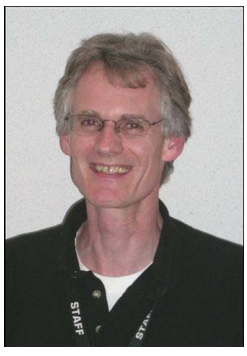
[6] Accellera UPF Standard version 1.0, February 2007, now IEEE standard 1891 http://www.accellera.org/activities/p1801_upf

[7] Si2 Common Power Format, CPF, specification http://www.si2.org/?page=811

[8] EUROPRACTICE mini@sic programme: http://www.europractice-ic.com/prototyping_minisic.php

[9] Kosonocky, S., "Practical Power Gating and DVFS", Hot Chips 23 Tutorial, Aug 2011 http://hotchips.org/uploads/hc23/HC23.17.1-tutorial1/Practical_PGandDV-Kosonocky-AMD.pdf

[10] Stan, M., "Low-Threshold CMOS Circuits with Low Standby Current," in Proceedings of the International Symposium on Low-Power Electronics and Design. Monterey, CA: IEEE/ACM, 1998, pp. 97–99



Dr David Flynn, a Fellow in R&D at ARM Ltd, has been with the company since 1991, specializing in System-on-Chip IP deployment and methodology. He holds a BSc in Computer Science from Hatfield Polytechnic, UK and a Doctorate in Electronic Engineering from Loughborough University, UK. He is currently part-time Visiting Professor with the Electronics and Computer Science Department at Southampton University, UK. David is a primary author of the Low Power Methodology Manual co-developed with Synopsys and launched in 2007.

High Performance State Retention with Power Gating applied to CPU subsystems - design approaches and silicon evaluation

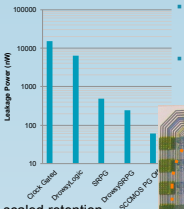
David Flynn, ARM - HotChips-24, August 2012

Building on basic Power Gating (as inferred by CPF, UPF) to address near-zero-overhead State Retention Power Gating, enhancing CPU leakage mitigation modes

- Power Gating, PG, is well supported by Electronic Design Automation tools
 - Based on logic-level drive of Header or Footer "MT-CMOS" switches
- State Retention Power Gating, SRPG, is also richly supported in the power intent standards

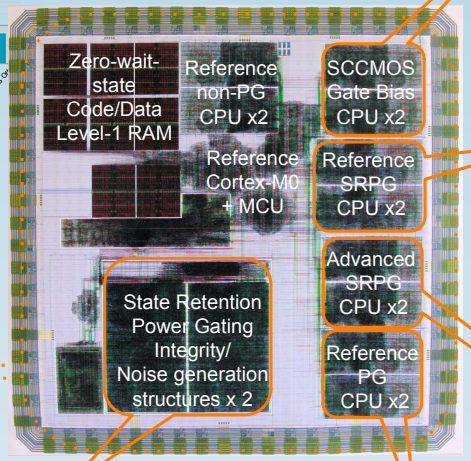
This work builds on industry standard EDA flows to allow more expert implementation techniques to be utilized without resorting to full-custom design

1. To support gate-bias (overdrive) techniques to enhance power gating
 - Super-Cutoff CMOS, SCCMOS, gate drive to power switches
 - Header power gates in this case
2. To provide advanced State Retention optimized for performance and minimal area impact
 - Advanced SRPG compared to the conventional EDA preferred register level abstraction
3. Combining the above to support additional leakage reduction modes including "drowsy" voltage scaled retention



Technology demonstrator

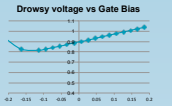
- Based on mature technology – TSMC65LP process "Tokachi-1" reference silicon
 - Including academic research (University of Southampton, UK)
 - With acknowledgement to EU EuroPractice "Mini-ASIC" research program
- Built using multiple instances of ARM Cortex®-M0 processor
 - 14 CPUs with one as a primary system MCU, 200MHz sign-off
 - Including 5 pairs optimized for performance (330MHz worst-case sign-off)
- Built using standard RTL and power intent
 - (Synopsys Inc. Multi-voltage EDA tools plus standard UPF)



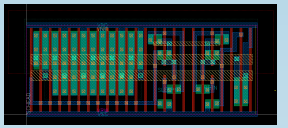
Super Cutoff power gating / Boosted-Gate drowsy retention

"Standard Cell" abstractions of

- Super-Cut-off Power Gating
 - Enhanced turn-off
 - Smaller Standard-Vt switches
- Boosted Gate drowsy retention
 - Very small Standard-Vt switches
 - Safer than full "diode-drop" Vret
- Clean EDA deployment





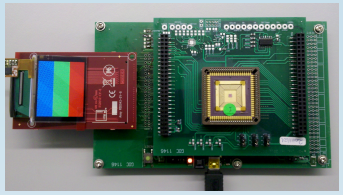




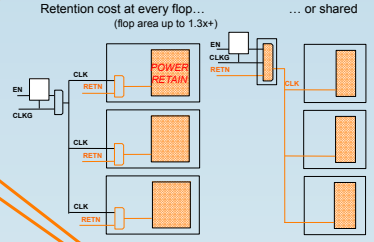
IP Deployment Reference designs

- Building on UPF and CPF PG inference
 - Well supported in Multi-Voltage tools
- Methodology and design flow "proving"
- Representative of real-world designs
 - In terms of multi-voltage challenges
 - But quick to design/validate/fabricate/analyze
- e.g. only 5M transistors, tiny 3.5mm^2 area
 - But 13 power domains, 3 VDD rails
 - Analog pads to observe "virtual" rails
 - Experimental structures to analyze integrity
- Artisan® libraries plus R&D prototype cells

Evaluation and characterization platform



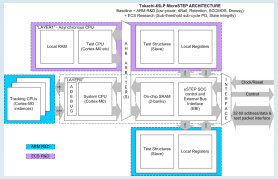
Addressing overheads of conventional SRPG



Low-impact SRPG

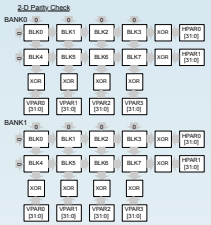
- Alternative to the EDA preference for Retention abstracted per-flop
 - With the associated area and performance cost
- "Clamp" (low) the clock from the final clock-gate
 - And insert a dummy clock gate where there is no gating
 - Share the cost of clamp across local cluster of flops
 - And "live with this" more expert "flow" in order to reap the benefits
- Reset also requires clamping
 - For asynchronous-reset flops
 - (without breaking test tools…)

SoC architecture



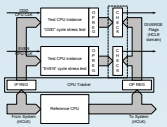
Retention integrity analysis structures

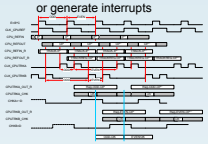
- Reconfigurable register (synthesized) arrays
 - Controlled noise / retention
- 2-D parity analysis
 - Bank0 or Bank 1
 - With noisy other bank
- Level-shift scan chains
 - To analyze VRET scaling
- First fail detect
 - Real-time compare
 - Level shifted detect
 - for voltage sensitivity
 - X-Y of first failing bit
 - Interrupt on error (raise voltage/wake/check)



Controlled overclocking harness

- MCU reference system uses pairs of experimental CPU macros
- Output comparators flag divergent behaviour
- Catches timing errors
- Can count cycles to error or generate interrupts

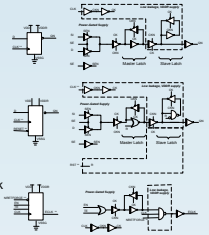




- Experimental macros clocked at double speed – 400MHz
- Fast-slow every other cycle to stay in-step with reference CPU
- Pairs required to ensure every instruction is run at-speed

HPSRPG Physical-IP abstraction

- "Live-Slave" flops
 - Dual-rail
 - Guarded clock
- Asynchronous Reset
 - As above
 - Guarded reset
- Clock Gate cells
 - Protect the clock



Prototyping the DySER Specialization Architecture with OpenSPARC

Jesse Benson, Ryan Cofell, Chris Frericks,
Venkatraman Govindaraju, Chen-Han Ho, Zachary Marzec, Tony Nowatzki,
Karu Sankaralingam
University of Wisconsin-Madison
Contact Email: karu@cs.wisc.edu

This paper describes the prototype implementation of the DySER specialization architecture integrated into the OpenSPARC processor. The paper's description covers the hardware, compiler, and application tuning. The prototype system provides speedups up to $14\times$ *over OpenSPARC (geometric mean* $5\times$*). The architecture is more flexible than SIMD and GPU- based acceleration while supporting a more diverse set of workloads.*

Overview Future processors must improve microarchitectural efficiency in order to overcome slowing transistor energy efficiency and sustain performance growth. The DySER architecture uses dynamic specialization to provide energy efficient performance improvements by complementing conventional processors. By using a co-designed hardware-compiler approach that avoids disruptive hardware or software changes, the architecture **Dy**namically **S**pecializes **E**xecution **R**esources to match application phases and achieves both functionality specialization (like Garp, Chimaera, Conservation-Cores) and parallelism specialization (like GPUs and SIMD short-vector extensions). We describe here the DySER architecture and its execution model, design and implementation of its compiler, prototype implementation, and conclude with performance results and significance of this work.

Architecture DySER is an array of configurable functional units connected with a circuit switched network of simple switches as shown in Figure 1. A functional unit can be configured to get its inputs from any of its neighboring switches. When all its inputs arrive, it performs the operation and delivers the output to a neighboring switch. Switches can be configured to route their inputs to any of their outputs, forming a circuit switched network. With this configurable network of functional units, a specialized hardware datapath can be created for a sequence of computation. To enable pipelining and dataflow like execution, both switches and functional units implement a simple credit based flow control that ensures data is forwarded only when the credit is available. Credits are generated when a functional unit/switch can accept new data. The switches in the edge of the array are connected to FIFOs, which are exposed to the processor core as DySER's input/output ports. DySER is tightly integrated with a general purpose processor pipeline, and acts as a long latency functional unit that has a direct datapath from the register file and from memory. The processor can send/receive data or load/store data to DySER directly through ISA extensions.

Execution Model Figure 2 shows DySER's execution model. Before a program uses DySER, it configures DySER by providing the configuration for functional units and switches. Then it sends data to DySER either from registers or from memory. Once data operands arrive at DySER's input FIFOs, they follow the configured path through the switches. When the data operands reach the functional units, the functional units perform the operation in dataflow fashion. Finally, the results of the computation are delivered to the output FIFOs, from where the processor fetches the outputs and sends them to the register file or to memory using ISA extensions. Further details are here [3, 2].

Compiler Design and Implementation DySER's compilation consists of four main phases and the key mechanism we leverage is the development of a new program representation called the Access-Execute Program Dependence Graph (AEPDG) that exposes the spatial and temporal aspects of dependences to the compiler. The four phases are : i) Selecting regions from the full program Program Dependence Graph (PDG) that are candidates for mapping to the DySER hardware. ii) Formation of the basic AEPDG encapsulating those code regions. iii) AEPDG transformation and optimizations to meet the goodness characteristics for the DySER architecture. iv) Code generation of the AEPDG. Our compiler implements a set of judiciously chosen and intuitive heuristics to produce good quality code as part of the transformations and optimizations phase. These are:

- Loop Unrolling/PDG Cloning
- Strip Mining/Vector Deepening
- Subgraph Matching
- Execute-PDG Splitting
- Scheduling Execute-PDG
- Loop Unrolling/Dependence Analysis
- Traditional Loop Vectorization
- Load/Store Coalescing.

To implement our compiler, we leverage the LLVM compiler framework and its intermediate representation(IR). We have developed LLVM optimization passes that process the LLVM-IR to construct the AEPDG and apply the associated transformations. Finally, we extend the LLVM code-generator to assemble DySER instructions and configurations.

Prototype Implementation We have completed a full RTL implementation of the DySER architecture integrated into the OpenSPARC pipeline. In terms of physical design, we have synthesis based results. The DySER block occupies an area of 1.54 mm^2 using a 55nm ASIC library, and on average consumes 72 mW.

In terms of implementation complexity, our prototype shows the DySER design is practical. The final interface consisted of only 11 signals in the RTL between OpenSPARC and DySER,

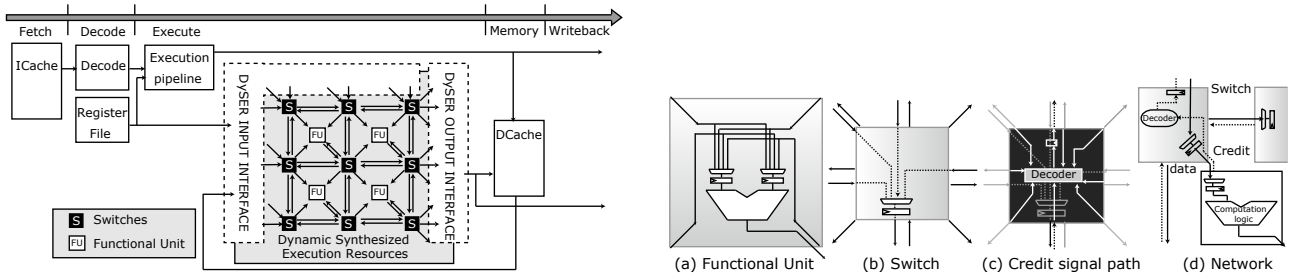


Figure 1: Processor Pipeline with DySER Datapath and DySER Elements

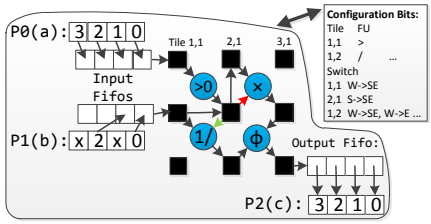
```
9

for(i=0; i<n; ++i) {
 if(a[i]>0)
   c[i] = 1/b[2i];
 else
   c[i] = b[2i]*2;
}
```

(a) Original Loop

```
9
DyConfig <Config>
for(i=0; i<n-n%4; i+=4){
  DyLd_Vec a[i:i+4]->P0;
  DyLd_Vec b[2i:2i+4]->P1;
  DyLd_Vec b[2i+4:2i+8]->P1;
  DySt_Vec P2->c[i];
}
...
```

(b) DySER-Accelerated Code



(c) DySER Configuration

Figure 2: DySER Execution Model

and a total of less than 750 lines of code modified in OpenSPARC. Further details are here [1].

We have also completed a mapping to FPGA of the full design using the Vertex-5 board. This FPGA implementation boots unmodified Ubuntu 7.10 Linux and runs C/C++ programs compiled through our toolchain. For detailed performance evaluation on our FPGA prototype, we implemented several FPGA optimizations to the architecture. These include simplifying the switches, "hardening" the configuration information and creating a FPGA bit-file specific to each application, and simplifications to the load-store interface.

Performance To measure DySER's efficiency in specialization, we have evaluated its performance on a suite of SIMD and GPU workloads to capture its functionality and parallelism specialization capability. We compare performance of DySER-accelerated implementations of these benchmarks to the sequential OpenSPARC implementation, and hand-optimized SIMD and GPU implementations. Based on measurements on our FPGA prototye implementation, compared to the OpenSPARC baseline, the DySER prototype, provides a speedup of up to 7×, with a geometric mean speedup of 3× on this diverse benchmark suite. Adding a vectorized mode to DySER provides up to 14× speedup with a geometric mean speedup of 5×. We observe that OpenSPARC's single-issue pipeline is the main bottleneck throttling the rate at which DySER is fed. When integrated with a dual-issue out-of-order processor, results from our *cycle-accurate performance simulator* show DySER continues to provide similar speedups: up to 14×, with a geometric mean speedup of 3.5×. As elaborated in [2], compared to SSE, DySER provides geometric mean 2.5× speedup, and compared to GPU execution, it provides 1.2× speedup.

Implications and Significance DySER is the culmination and generalization of trends already occurring for popular parallelism based accelerators. SSE has been augmented with both functionality-specialized and non-purely word parallel instructions. Instructions in NVIDIA Kepler GPUs are specialized for the particular region with compiler annotations indicating when to issue.

Not only is DySER a more natural evolution of specialization strategies, but it is also more practical to implement. From a software perspective, it is a more flexible compiler target than SSE, and DySER does not require a new software stack and application implementations as for the GPU. From a hardware perspective, its interface enables simple integration with a processor pipeline.

The most profound implication of DySER is that the execution model and architecture provide a practical way to implement instruction-set specialization, SIMD specialization, and domain-driven accelerators using one substrate. With its impressive speedup and corresponding energy gains, DySER significantly improves architectural energy efficiency using specialization. The novel architecture, its prototype implementation, and energy efficiency implications of the execution model provide a set of promising mechanisms.

References

[1] J. Benson, R. Cofell, C. Frericks, C.-H. Ho, V. Govindaraju, T. Nowatzki, and K. Sankaralingam. Design Integration and Implementation of the DySER Hardware Accelerator into OpenSPARC. In *HPCA '12*.

[2] V. Govindaraju, C.-H. Ho, T. Nowatzki, J. Chhugani, N. Satish, K. Sankaralingam, and C. Kim. DySER: Unifying Functionality and Parallelism Specialization for Energy Efficient Computing. *IEEE Micro*, 32(5), 2012.

[3] V. Govindaraju, C.-H. Ho, and K. Sankaralingam. Dynamically Specialized Datapaths for Energy Efficient Computing. In *HPCA '11*.

Prototyping the DySER Specialization Architecture with OpenSPARC

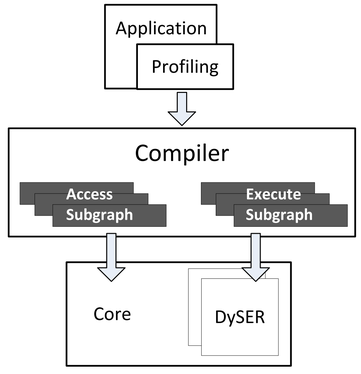
Jesse Benson, Ryan Cofell, Chris Frericks, Venkatraman Govindaraju,
Chen-Han Ho, Zachary Marzec, Tony Nowatzki, and Karu Sankaralingam

VERTICAL

THE UNIVERSITY of WISCONSIN MADISON

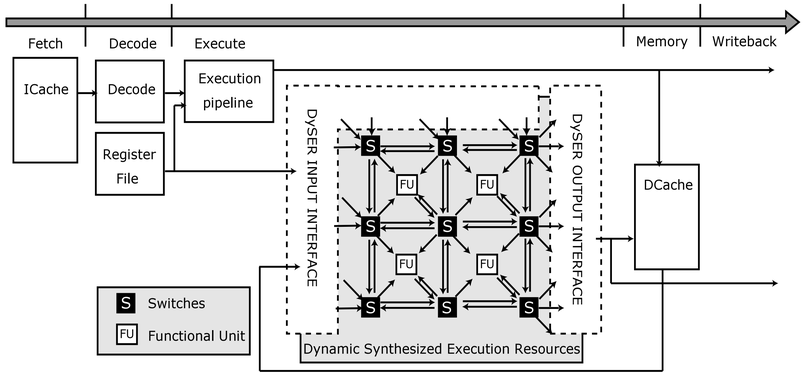
DySER Approach

- Compiler assisted dynamically specialized computation through heterogeneous array of functional units
- DySER configured once for multiple invocations



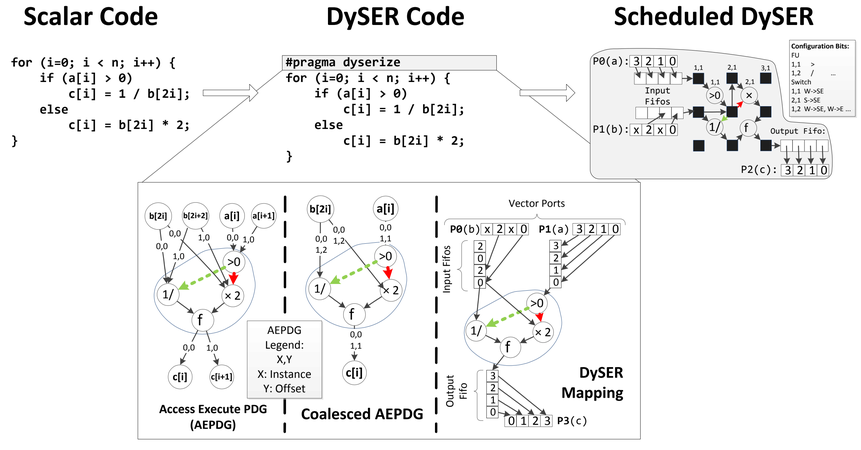
DySER Architecture

- Alongside Execute stage in processor pipeline
- Concurrently executes DySER and non-DySER code



DySER Compiler

- LLVM based compiler
- Generates specialized binaries for DySER from C/C++ source code



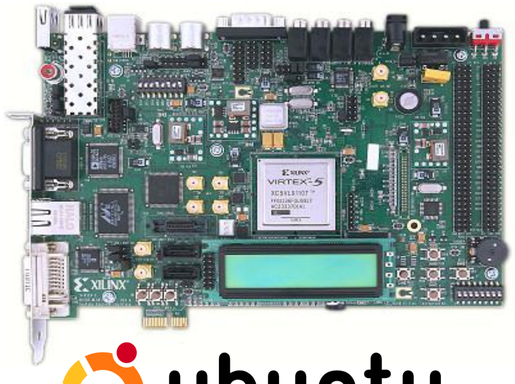
OpenSPARC T1 Integration

- Limited ISA extensions required
 `d_init, d_send, d_recv, d_load, d_store`
- Only **eleven** interface signals in RTL/microarch
- Few lines of changed Verilog code:

Unit	Lines Changed	Notes
IFU	275	Reserved opcodes used for DySER Instructions
LSU	23	Reverse engineered memory control
EXU	216	DySER model Verilog and 18 FF added
MMU	0	Unchanged!
Total	514	Minimal changes!

FPGA Prototype

- Utilizes Xilinx Virtex5 FPGA Board
- Fits a “hard” 4x4 DySER with fixed paths
- Boots unmodified OpenSPARC Ubuntu 7.10
- DySER is not on the critical path!



ubuntu

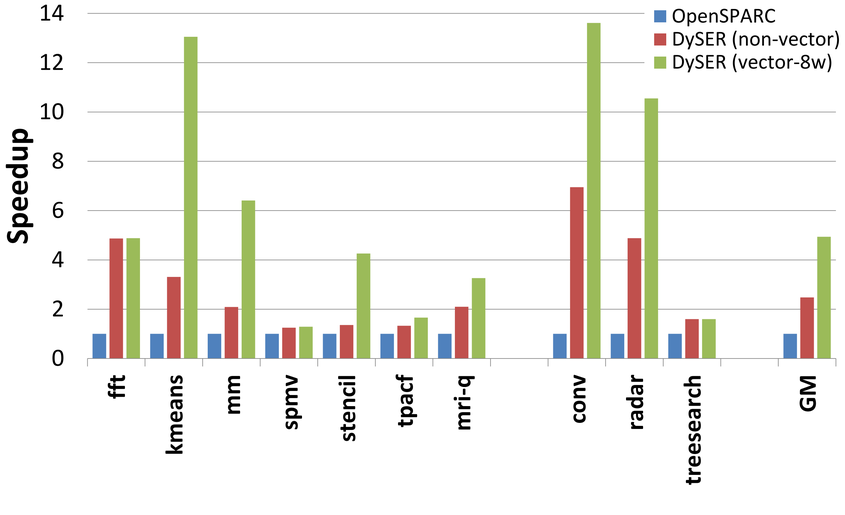
ASIC Synthesis @ 55nm

Area: 1.54mm^2 Power: 72mW

DySER Performance

- Throughput/high-performance workloads
- Competitive or surpasses SIMD/GPU approach

Non-vectorized (1 wide) : 2.5x speedup
Vectorized (8 wide): 4.9x speedup

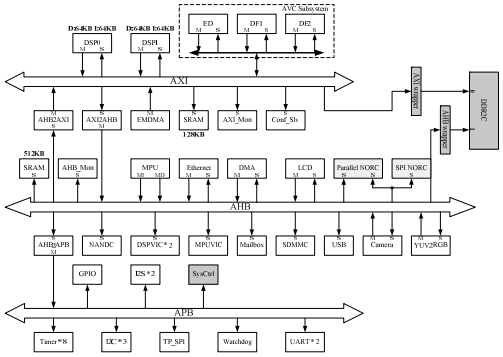


Low Power and High Performance 3-D Multimedia Platform

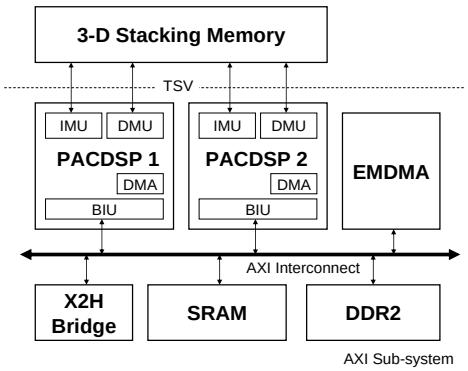
Po-Han Huang, Chi-Hung Lin, Hsien-Ching Hsieh, Huang-Lun Lin and Shing-Wu Tung
Information and Communications Research Lab.
Industrial Technology Research Institute
Hsinchu, Taiwan
E-mail: pohan@itri.org.tw

Traditional technology scaling of semiconductor chips followed Moore's Law. However, the transistor performance improvement will be limited, and designers will not see doubling of operating frequency every two years. Recently, 3-D integrated circuits that utilize through silicon via (TSV) for interconnection have been developed as an improved alternative to the Package-on-Package (PoP) and System-in-Package (SiP) packages. There are many benefits by using TSV-based 3-D integration technologies: (1) Circuit delay can be improved due to the shorter interconnect and reduced parasitic capacitance/inductance, (2) more functionality can be integrated into a small silicon space for form factor reduction and higher packing density due to the additional third dimension, (3) different components with incompatible manufacturing process (i.e. Logic, DRAM, Flash, etc) can be combined in a single 3-D IC for heterogeneous integration. The 3-D integration based on TSV technology enables stacking of multiple memory layers to obtain higher bandwidth for the recent multimedia applications at lower energy consumption. Intel has demonstrated through the teraflops microprocessor chip which is an 80-core design with memory-on-logic architecture. And, each core connects to a 256KB SRAM with 12GB/s bandwidth. Although 3-D IC overcomes many limitations and drawbacks on 2-D IC design, it still has many challenges and design issues that should be considered carefully. In general, the number of TSV is the most critical constraint while designing a 3-D architecture because it is highly related to system performance.

This poster presents a 3-D multimedia platform – 3D-PAC designed by ITRI. 3D-PAC is developed by stacking original 2D-PAC with the SRAM tier. Based on this 3-D stacking technology, the performance can be enhanced about nearly 54% according to different applications. And this poster will show the method of architecture exploration for 3-D stacking. It also describes the detail implementation of reconfigurable SRAM and tier selection when multi-layer stacking SRAM is needed. Finally, the chip is fabricated in TSMC 90nmG CMOS technology. 2D-PAC has the novel features which are described as follows. It is a heterogeneous multi-core architecture, composed of an ARM926EJ-S and two PACDSPs (variable-length, 5-way VLIW architecture designed by ITRI/ICL). This system also consists of three different kinds of buses: AXI, AHB and APB. There are also many peripherals implemented in the system such as I^2C, UART, …, etc.



In general, the new architecture evaluation for an optimized stacking static memory is driven by area, performance, energy efficiency, number of TSVs and thermal issues. After architecture exploration using the electronic system level (ESL) design, the stacking memory is integrated with the instruction memory unit (IMU) and data memory unit (DMU) of PACDSP core because of the performance and number of TSV to build up so called 3D-PAC platform.



For this architecture, each PACDSP owns its private SRAM block (256KB) which is reconfigurable. It allows programmers to configure the different architectures of stacking memories for different applications. For example, a user can configure a part of memory as instruction memory and the rest as data memory or make the entire memory as data memory, which provides high flexibility for the original architecture. With the 3-D stacking SRAM, it equally extends the instruction cache and data memory size of PACDSP. That means programmers can profit in two ways: (1) Reduce the latency caused by cache misses with increasing the size of instruction cache, (2) reduce the frequency of external memory accesses with increasing the size of local data memory.

Two real applications, multi-channel H.264 decoder and JPEG decoder, are chosen to analyze the impact and efficiency from extending PACDSP local memory by 3-D stacking. Experiments are executed on the ESL platform mentioned before. In Multi-channel H.264 decoder application, it takes two PACDSPs to decode four different films simultaneously and display on LCD screen at the same time. It requires more data movement and computation power compared to the

single film decoding.

Experiment configurations (H.264 decoder):

- ARM9 = 204MHz, AHB = 102MHz
- PACDSP = 204MHz, AXI = 204MHz
- DDR2 data rate = 408MHz
- Bitstream:
 SHISEIDO_track1,
 Jomo_track2,
 Ice_Age,
 STC_TEST_Motion
 (10 frames, QVGA)

Experimental results show that overall performance can improve from 11.56fps to 14.80fps (28.02%) without applying any parallelism and optimization for H.264 decoder application. Each DSP is in charge of the decoding of two films. Without 3-D stacking memory, PACDSP needs to backup the internal data to external DDR2 memory during the film switching because of lack of internal data memory, and it incurs a huge overhead. By contrast, with enough 3-D stacking memory supporting (each 256KB), PACDSP does not have to backup related data during film switching. So that will make a huge improvement for system performance. After applying some parallelism and optimization to multi-channel H.264 decoder, the system performance can reach 26.09fps (54.19% improvement compared with 2D 16.92fps).



H.264 Version	Total Cycle	FPS	Improve.
Non-Parallel Version			
2D	194,019,671	11.56	-
3D with 4 binary	151,524,469	14.80	28.02%
Parallel Version			
2D	164,531,751	13.64	-
3D with 2 binary	121,891,167	18.40	34.89%
Parallel Version (Enhanced)			
2D	132,557,335	16.92	-
3D with 2 binary	86,001,705	26.09	54.19%

Experiment configurations (JPEG decoder):

- ARM9 = 204MHz, AHB = 102MHz
- PACDSP = 204MHz, AXI = 204MHz
- DDR2 data rate = 408MHz
- Bitstream: test_image (1 frame, QCIF)

Experiment result shows that there is only little system improvement by extending the instruction cache for JPEG decoder because of the low ratio of cache misses (0.02%). It stands that the original 64KB cache is enough for this application. By contrast, there is huge performance improvement by enlarging the local data memory for JPEG decoder because of the high ratio of external accesses (85.87%). It means the original 64KB data memory is not enough for this application. According to this analysis, it seems that suitable memory configuration depends on different applications. 3D-PAC platform maintains this design feature. Each PACDSP owns its private SRAM block (256KB) which is reconfigurable. It allows programmers to configure the different architectures of stacking memories for different applications.

Execution Type	Cycle Count	Ratio
Cache Size : 64KB		
Cache miss	129,314	0.02%
External access	550,498,586	85.84%
DSP computation	90,676,349	14.13%
Total cycle count		641,304,249
Cache Size : 128KB		
Cache miss	9,376	0.00%
External access	550,494,242	85.87%
DSP computation	90,525,505	14.12%
Total cycle count		641,029,123

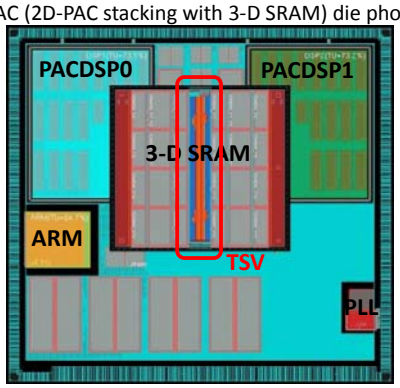
Execution Type	Cycle Count	Ratio
Data Memory Size : 64KB		
Cache miss	9,376	0.00%
External access	550,494,242	85.87%
DSP computation	90,525,505	14.12%
Total cycle count		641,029,123
Data Memory Size : 64+192KB		
Cache miss	8,865	0.00%
External access	2,051,198	1.96%
DSP computation	102,449,710	98.02%
Total cycle count		104,509,773

There are total 1,914 TSVs in 2D-PAC allocated in the middle area. Related chip SPEC (both 2D-PAC and 3-D stacking SRAM) and die photo shows as follows:

Design	2D-PAC
Process	TSMC 90nmG
Operating Frequency	PACDSP 300MHz
Operating Voltage	Core: 1.0~1.2V I/O: 3.3V
# I/O Pads	498 with PWR/GND
Die Area	7880 x 7880 um^2

Design	3-D stacking SRAM
Process	TSMC 90nmG
Operating Frequency	300MHz
Operating Voltage	Core: 1.0~1.2V
Die Area	3880 x 3880 um^2

3D-PAC (2D-PAC stacking with 3-D SRAM) die photo:



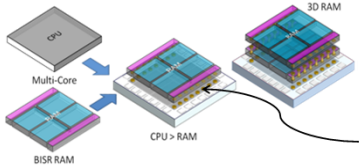
Low-Power and High-Performance 3-D Stacking Multimedia Platform

Industrial Technology Research Institute (ITRI), Taiwan

Po-Han Huang
Chi-Hung Lin
Hsien-Ching Hsieh
Huang-Lun Lin
Shing-Wu Tung

- A low-power & high-performance 3-D stacking multimedia platform is proposed
 - Heterogeneous (logic & memory) integration via 3-D technology
 - Extensive architecture exploration with ESL simulation
- 54.19% performance improvement over 2-D architecture
- Reconfigurable stacking SRAM improves the programming flexibility, where the performance can be optimized for different applications
 - Overhead is neglectable
- Fabricated in the TSMC 90nm generic CMOS technology
 - 1,914 TSVs have been utilized for 3D stacking

Integration of 3D stacking technology for 3D-PAC

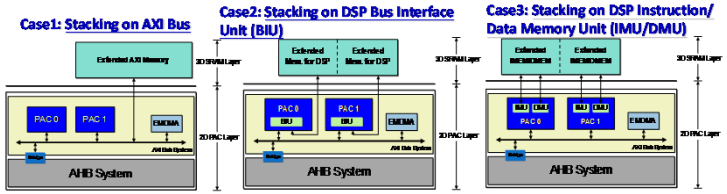


3D-PAC for system bandwidth and performance improvement

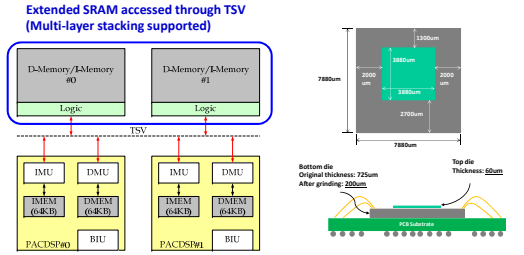
Tradeoff

Combined with TSV

Architecture exploration for 3D-PAC



Final architecture of 3D-PAC



Design SPEC:

- ☐ Private SRAM block (256KB) for each DSP
- ☐ Reconfigurable SRAM as internal instruction or data memory of DSP

Implementation Results

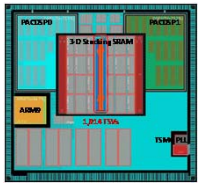
Logic Layer

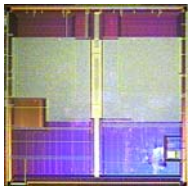
Design	2D-PAC
Process	TSMC 90nmG
Operating Frequency	PACDSP 300MHz
Operating Voltage	Core: 1.0~1.2V I/O: 3.3V
# I/O Pads	498 with PWR/GND
Die Area	7880 x 7880 um²

Memory Layer

Design	3-D stacking SRAM
Process	TSMC 90nmG
Operating Frequency	300MHz
Operating Voltage	Core: 1.0~1.2V
Die Area	3880 x 3880 um²

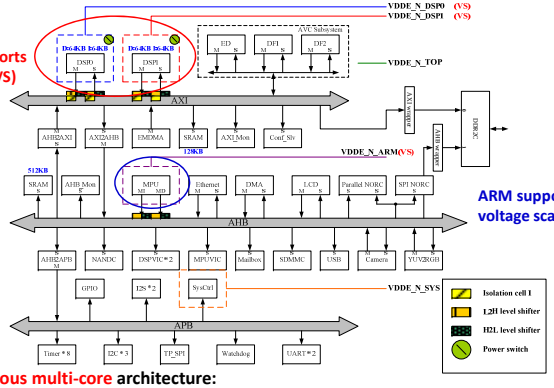
Heterogeneous integration with total 1,914 TSVs







2D-PAC with various low-power techniques



Both DSPs can be turned off or supports voltage scaling (DVS)

ARM supports voltage scaling (DVS)

Heterogeneous multi-core architecture:

- ☐ ARM926EJ-S
- ☐ Two PACDSPs (variable-length 5-way VLIW)

Low power architecture design flow:
Common Power Format (CPF) by Cadence Inc.

Performance/Cost Evaluation with ESL

Configurations:

- ☐ ARM9 = 204MHz, AHB = 102MHz
- ☐ PACDSP = 204MHz, AXI = 204MHz
- ☐ DDR2 data rate = 408MHz
- ☐ Application: H.264 decoder
- ☐ Bitstream: foreman (30 frames, QCIF)

Architecture	Total Cycle	FPS	Improve.	# TSV
2D-PAC (DDR2)	164,531,751	13.64	-	-
3D-PAC (3-D SRAM on AXI bus)	162,312,919	13.83	1.39%	272
3D-PAC (3-D SRAM on BIU)	162,308,407	13.83	1.39%	544
3D-PAC (3-D SRAM on IMU/DMU)	121,891,167	18.40	34.89%	1,886

✓

Configurations:

- ☐ ARM9 = 204MHz, AHB = 102MHz
- ☐ PACDSP = 204MHz, AXI = 204MHz
- ☐ DDR2 data rate = 408MHz
- ☐ Application: JPEG decoder
- ☐ Bitstream: test_image (1 frame, QCIF)

Case1: Cache size

Execution Type	Cycle Count	Ratio
Cache Size: 64KB		
Cache miss	129,314	0.02%
External access	550,498,586	85.84%
DSP computation	90,676,349	14.13%
Total cycle count	641,304,249	
Cache Size: 128KB		
Cache miss	9,376	0.00%
External access	550,494,242	85.87%
DSP computation	90,525,505	14.12%
Total cycle count	641,029,123	

Case2: Data memory size

Execution Type	Cycle Count	Ratio
Data Memory Size: 64KB		
Cache miss	9,376	0.00%
External access	550,494,242	85.87%
DSP computation	90,525,505	14.12%
Total cycle count	641,029,123	
Data Memory Size: 64+192KB		
Cache miss	8,865	0.00%
External access	2,051,198	1.96%
DSP computation	102,449,710	98.02%
Total cycle count	104,509,773	

Suitable memory configuration depends on different application => Reconfigurable 3-D stacking memory

ITRI Industrial Technology Research Institute

The Model Is Not Enough: Understanding Energy Consumption in Mobile Devices

James Bornholt
Australian National University
u4842199@anu.edu.au

Todd Mytkowicz
Microsoft Research
toddm@microsoft.com

Kathryn S. McKinley
Microsoft Research
mckinley@microsoft.com

1. Introduction

Although battery life has always constrained embedded and mobile hardware developers, the rise of smart phones and tablets has also made energy a fundamental concern of software developers. On the desktop, software developers generally ignored energy, but in the mobile environment, battery life is critical to the user experience. Just as developers use performance profiling tools, they now need energy profiling tools to understand how and why their software consumes energy.

The inconvenience, cost, and complexity of external power measurement hardware and the inaccuracy of on-board power sensors on older phones [1] motivated researchers to create power modelling tools. Power modelling uses utilization metrics to estimate power draw based on previously measured correlations between the metrics and power.

We show that the on-board power sensor is now accurate on a Windows Phone 7.5 device running on a SnapDragon MSM8660. Compared to external measurement hardware, the on-board "fuel gauge" is accurate to within 2% of total energy consumption. We thus modify the Windows Phone 7.5 OS to sample power without external hardware, sample the application call stack to correlate energy consumption with code, and examine power traces from two weeks of normal use. These traces illustrate behavior where modelling alone is not sufficient to understand the energy consumption of a mobile device. For example, we observe inter-day variations in base power draw as the battery discharges, an effect that to our knowledge is not captured by existing modelling work.

This work recommends that a hybrid approach will improve the accuracy of energy profiles, and that direct measurements will significantly improve the accessibility of fine-grained energy information in both testing and deployment. Armed with easy-to-use energy analysis tools, hardware designers, OS developers, and third party application developers will be better equipped to understand and optimize the energy behavior of mobile code.

2. Related Work

Early energy modelling research used power measurements of executing each machine instruction [4, 5]. These methods do not extend well to modelling the power draw of other non-CPU components, which constitute two thirds or more of energy consumption on mobile devices.

For mobile devices, recent models use linear regression trained on energy profiles of the entire device gathered from scenarios that stress each device component [3]. This model is accurate compared to external measurements on short-duration test runs. This approach, however, cannot address the tail power state problem, where components improve responsiveness by waiting in a higher power state for more work to arrive (see Section 4.1).

To provide fine-grained energy accounting, Pathak et al. [2] introduce a finite state machine that models component power draw by tracing system calls. For example, a `read` system call transitions the flash storage component into a higher power state. This approach is promising for attributing energy consumption to code. However, it does not model power draw of key device components, such as the screen or application CPU draw, and so cannot present a complete picture of the device's energy usage.

Dong and Zhong [1] overcome these problems with a hybrid approach. They create models on-the-fly using the on-board battery sensor. They use low-frequency samples from the sensor to track accuracy and trigger model reconstruction if the variation is too high. This method potentially accounts for variation in base power draw, but with a significant cost or latency. However, it does not address the tail power state problem.

3. Measuring energy consumption

External measurement hardware, such as the Power Monitor with which we compare in this paper, accurately measures power draw from a phone's battery. These tools, however, are relatively expensive ($750, which is more than the phone itself) and limit phone mobility, restricting real world testing.

On modern mobile devices, the "fuel gauge" (FG) provides accurate readings of battery voltage and instantaneous current draw for displaying remaining battery life. To validate the FG's accuracy, we executed benchmarks on a HTC (MSM8660 processor) device running Windows Phone 7.5, simultaneously capturing power measurements from the FG and an external Power Monitor. The FG was accurate to within 2%$\pm$0.02 of the Power Monitor. We suggest this accuracy is enough to replace external hardware as the source of power measurements.

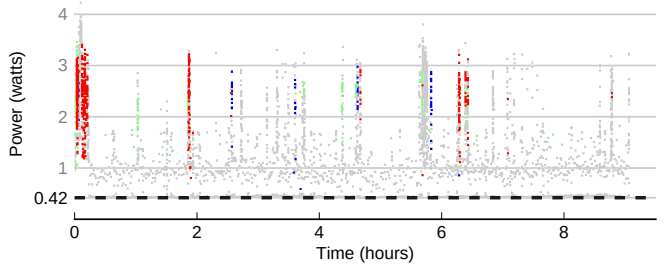
4. Power modelling

The main challenge in on-board energy profiling is accurately attributing the energy consumed by particular applications and methods. The traditional approach to this problem has been *modelling*, which can both produce power readings and attribute them to code entities.

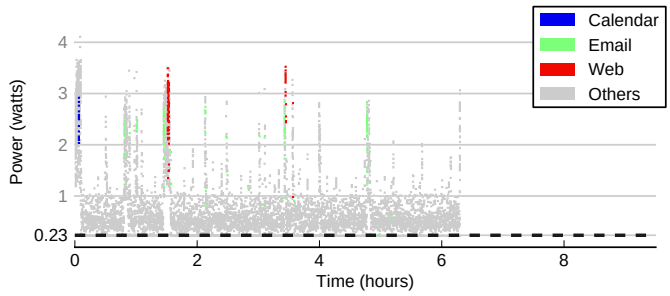
4.1 Tail power states

Even with a model, *tail power states* complicate energy profiling. To provide responsiveness, many components (e.g., radio, GPS) continue to draw high power after use. For example, a 3G radio may remain in a higher "tail" power state for up to seven seconds after use. This tail power state complicates energy attribution: the download has completed, the code has moved on, and the application may no longer be running, but power is still drawn. Further, if several applications use a component, which one should be charged for the tail power state?

Pathak et al. [2] model tail power states with their system call model, record the calling context of each system call, and assign the tail power to the last calling context that used the device.



(a) Day 1: Idle power 0.42 W



(b) Day 2: Idle power 0.23 W

Figure 1. Two days of power usage with day-to-day variation. Points are coloured by active application.

4.2 Shortcomings of modelling

There are two main drawbacks to modelling:

1. Accuracy is limited by the training environment and components modelled (e.g., CPU and screen are often missing). Because the mobile environment is so diverse (hardware models, cellular networks, etc.), modelling has the potential to be very inaccurate.
2. The trade-off between latency, cost, and portability essentially limits models to testing. Shipping system call traces to a server introduces long latency, but running models on-board has high overhead and can require external hardware, impeding portability.

Our experiments also show a significant day-to-day variation in base power draw when the phone is "idle." Figure 1 plots two days of power readings, revealing a variation of 0.19 W in base power draw between two consecutive days of usage (the dashed line at base of each figure). A point (x, y) on this graph plots the power in watts (y) as a function of a day's usage (time on the x axis). Day 1's base draw is 0.42 W while Day 2's is 0.23 W, which is a significant difference in total energy consumption over several hours. Unless this variation is observed in the lab, a model is unlikely to predict this day-to-day variation.

5. A hybrid approach

We advocate a hybrid approach to accurately attribute energy consumption to the code that uses it: accurate battery sensors (like the FG) to gather online power readings, and system call modelling to handle components with tail states.

This approach has several benefits. First, it ensures that an energy profile always shows a complete picture of system energy consumption. By always capturing whole system power readings, components acting in ways unanticipated by the model do not go unobserved. While precise attribution is challenging, the recent applications and their call stacks will help identify problems, particularly if applications are sampled over many days. Real measurements at the very least identify power anomalies, and that the model may need correction. Second, a hybrid approach deals with tail power states in a way measurement alone cannot. For example, it can accurately assign the energy usage of networking hardware to the code making network calls, providing a more actionable energy profile to a developer. Finally, if we use on-board sensors, we can avoid the need for expensive external hardware, improving the accessibility of energy information to developers.

This approach also opens a new field of potential optimization at the OS level. With completely online power measurements, the OS may monitor battery life performance at a fine-grain in real time, and perform optimization to achieve a battery life goal. For example, if an alarm is set for 8 hours in the future and projecting current energy consumption indicates that the battery will not last that long, the OS can optimize components to meet this power goal. This approach improves over the current practice, in which the OS takes drastic measures when the battery is very low and it is too late to recover. Guided by the component attribution possible with modelling, the OS may determine that the GPS is consuming too much energy and tune it to use less energy by providing less accurate positions. Gathering the data must have low overhead (less than 4% in our testing), such that it does not contribute to the problem. The OS needs fine-grained real-time low cost power data to make these types of optimizations. Our tool provides such data.

6. Conclusion

Mobile devices are placing energy efficiency in the hands of software developers. Unfortunately, power modelling alone cannot identify significant power variations, nor is it practical in deployment. We show that on-board fine-grained power measurements of the fuel gauge are now both accurate and low overhead.

We advocate a hybrid approach to power measurement: combining the best aspects of both modelling and measurement to produce accurate and actionable energy information for developers. Using such tools, developers have the potential to understand and optimize the energy behavior of their code. Operating system developers have the potential to guide real-time optimization in the interests of battery life, a significant advancement over the current state of power optimization. These advances are critical to the future of mobile software, as developers at all levels come to terms with their new-found responsibility for energy efficiency.

References

[1] M. Dong and L. Zhong. Self-constructive high-rate system energy modeling for battery-powered mobile systems. In *Proceedings of the 9th international conference on Mobile systems, applications, and services*, MobiSys '11, pages 335–348, June 2011.

[2] A. Pathak, Y. C. Hu, M. Zhang, P. Bahl, and Y.-M. Wang. Fine-grained power modeling for smartphones using system call tracing. In *Proceedings of the sixth conference on Computer systems*, EuroSys '11, pages 153–168, Apr. 2011.

[3] A. Shye, B. Scholbrock, and G. Memik. Into the wild: studying real user activity patterns to guide power optimizations for mobile architectures. In *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO 42, pages 168–178, Dec. 2009.

[4] P. Stanley-Marbell and M. Hsiao. Fast, flexible, cycle-accurate energy estimation. In *Proceedings of the 2001 international symposium on Low power electronics and design*, ISLPED '01, pages 141–146, 2001.

[5] V. Tiwari, S. Malik, A. Wolfe, and M. T.-C. Lee. Instruction level power analysis and optimization of software. In *Proceedings of the 9th International Conference on VLSI Design: VLSI in Mobile Communication*, VLSID '96, pages 326–328, 1996.

The Model Is Not Enough: Understanding Energy Consumption in Mobile Devices

James Bornholt, Australian National University, u4842199@anu.edu.au
Todd Mytkowicz, Microsoft Research, toddm@microsoft.com
Kathryn S. McKinley, Microsoft Research, mckinley@microsoft.com

Why understand energy?

Smart phones and tablets force software developers to focus on battery life.

Developers already optimise performance using profilers. Our goal is to build an energy profiler.

Why is energy profiling difficult?

The two ways to calculate energy—hardware power meters and software modelling—have drawbacks.

Attribution of energy to code is made difficult by tail power states, which shift the blame.

Tail power states: who's to blame?

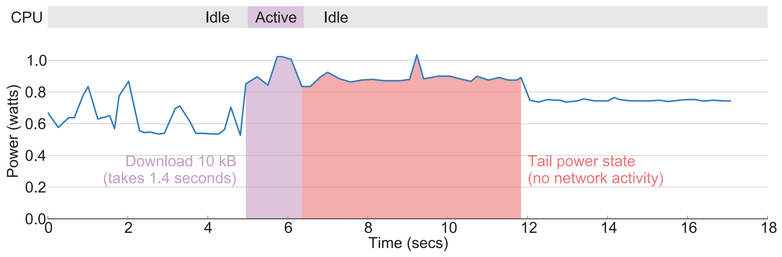


Figure 1 The 3G radio's tail power state consumes energy while the CPU is idle.

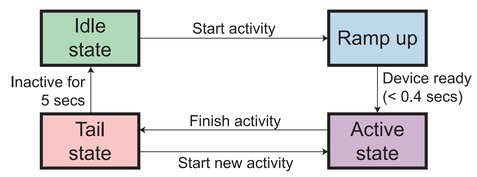


Figure 2 Tail power states avoid returning to the ramp up state, therefore reducing latency.

A deeper look: energy profiling

Why understand energy?

- Software and energy interact in subtle and non-trivial ways – for example, grouping of network requests to avoid thrashing the network hardware
- Energy bugs present serious usability problems, because users cannot easily identify them until it is too late (i.e. the battery is empty)

Why is energy profiling difficult?

- Hardware meters are often expensive and bulky
- Models are specific to their training environment
- Tail power states reduce ramp-up latency by allowing components to remain powered on after last use
- So it's wrong to attribute current power draw to the currently executing code – the real culprit may be long gone!

Power modelling

Modelling uses metrics such as CPU and network usage to extrapolate power draw, but this has issues for profiling use.

Recent work uses system call tracing to handle tail power states, but other issues remain.

Power measurement

Hardware power meters are common for power measurement, but these are impractical for most developers.

Our work shows the onboard "fuel gauge" battery sensor is accurate to within 2% of external meters, but measurement alone cannot address tail power states.

Modelling misses important variation

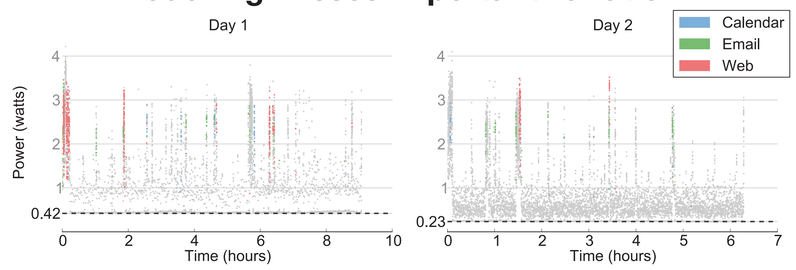


Figure 3 Modelling misses a base power variation, worth 30% of battery life over 8 hours.

A deeper look: modelling and measurement

Power modelling

- The raw numbers are often quite good: errors of below 10%
- But the models:
 1. Do not address tail power states
 2. Are specific to the lab environment they are trained in
 3. Cannot detect the unexplained power variation in Figure 3
- In system call tracing, finite state machines model each component's power states, using system calls to transition the machine
- This addresses the tail power state problem, but suffers the other drawbacks of modelling, and also cannot yet capture important components like the screen

Power measurement

- External meters are expensive and bulky, making them inaccessible and difficult to use
- The fuel gauge typically measures battery capacity, but modern sensors provide instantaneous power draw
- When testing an HTC Windows Phone, the onboard sensor was accurate to within 2% of total energy compared to the power meter
- The overhead of sampling the fuel gauge at 5 Hz was less than 4%.
- This means the fuel gauge is accurate enough to replace the external power meter for many uses, but may not be appropriate for very high frequency sampling
 - The power meter we used samples at 5000 Hz

A hybrid approach

Based on our results, we advocate a hybrid approach to energy profiling.

This approach makes energy profiling more accurate, more actionable, and more accessible.

What do we get from this data?

Energy profiles let developers isolate poor energy usage in their code.

Energy profiles open a new field of potential online OS-level energy optimisations.

Energy profiles identify real bugs

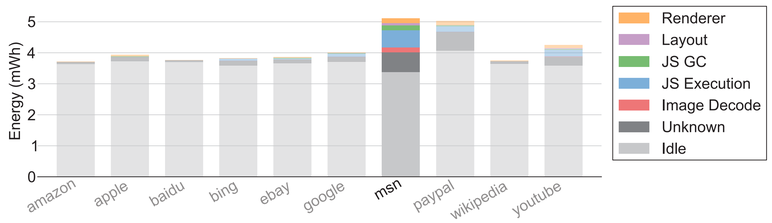


Figure 4 An energy bug in the MSN website, seen and diagnosed by an energy profiler.

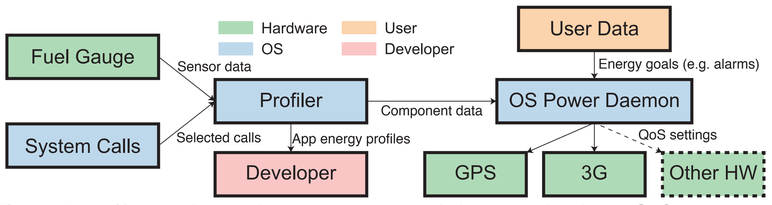


Figure 5 Profiles can be used to meet energy goals by tuning hardware QoS settings.

A deeper look: hybrid energy profiles

A hybrid approach

- Accurate onboard sensors for online, whole-system power readings, and system call modelling to handle tail power states
- Energy profiles will always show a complete picture of system energy use, rather than omitting unmodelled components, making profiles more accurate
- Tail power states are handled, making profiles more actionable
- By removing the need for hardware, even for training, profiles are more accessible

What do we get from this data?

- In Figure 4 we see MSN as a clear outlier in energy use
- The profile identifies excessive image preloading as the cause of this inefficiency
- The OS can use online profiles to achieve battery life goals; for example, tuning energy use to ensure a scheduled alarm goes off
- Guided by component attribution, this tuning can target specific energy users like the GPS, and trade accuracy or speed for battery life

Australian National University

Microsoft Research

Efficient, Precise-Restartable Program Execution on Future Multicores

Gagan Gupta, Srinath Sridharan, and Gurindar S. Sohi
Department of Computer Sciences, University of Wisconsin-Madison
{gagang, sridhara, sohi}@cs.wisc.edu

1. INTRODUCTION

Multicore processors are becoming ubiquitous, placing new demands on hardware and software designers. No longer do a small set of experts develop a few software applications for a small number of parallel machines. Already standard in servers, desktops and laptops, today handheld devices use multicores, expanding the spectrum of their use from mobile computing at the low end to cloud computing at the high end. Consequently, dramatically increased number of software developers are creating hundreds of thousands of applications to run on a plethora of diverse platforms. Thus ease of writing parallel programs, to achieve energy and/or performance efficiency, continues to gain importance.

At the same time, programmers have to account for the changing characteristics of emerging technologies. Processors are transitioning from homogeneous cores to heterogeneous cores with disparate performance/energy characteristics. As future computing hardware pushes the limits of semiconductor technology, it will become increasingly unreliable. Simultaneously, emerging use of computing systems will require them to host multiple applications concurrently, even on mobile devices. Unreliability, resource (computing and energy) management, and service-level agreements will lead to imprecise knowledge of available resources during a program's execution. Hence programmers can no longer assume availability of given (or constant) resources to process an application, unlike in canonical parallel programming.

The confluence of the above factors pose daunting challenges to programmers in writing ubiquitous programs and achieving their reliable, energy-efficient, parallel execution, while remaining agnostic of the unpredictable, dynamically (and potentially continuously) changing computing conditions.

We propose a model that seamlessly addresses this range of challenges. It relies on expressing parallel algorithms as sequential programs, i.e., *statically-sequential* (§2.1), and performing their controlled, dynamic parallel execution while honoring their sequential semantics. Although at first glance the approach may appear antithetical to parallelism, we show that it affords several advantages. Its intuitive interface and sequentially determinate execution (which ensures that in any execution of a program with the same inputs, a variable is assigned the same sequence of values) allow programmers to easily reason about the program execution, simplifying programming. The model utilizes the implied order in a statically-sequential program to achieve a dataflow schedule of parallel execution (§2.2), potentially exploiting all available parallelism. Further, the order permits the adaptability needed to achieve efficient execution in dynamically changing (§2.3), unreliable (§2.4) computing environments. We provide an overview of these aspects and present results from our efforts to develop several benchmark applications using the model, implemented as a fully functional runtime system, on stock multicore systems.

2. DYNAMIC PARALLELIZATION OF SEQUENTIAL PROGRAMS

Our approach strives to minimize the burden on programmers. It allows programs to be authored in established imperative programming languages, such as C++, and automates their parallel execution. The model extracts a program's computations, establishes the dynamic data-flow between them, and schedules their ordered execution as the prevailing resources permit. It can also roll back the execution, up to a desired point, and resume it, if desired. We highight the model's principles by describing the programming interface and the mechanisms as implemented in the runtime (a C++ library).

2.1 Composing Programs

Programmers today follow modern software engineering and object-oriented (OO) design principles by composing programs from reusable functions that manipulate encapsulated data and communicate with each other using well-defined interfaces. Often such "well-composed" functions avoid side-effects by only manipulating data communicated through the interfaces. We seek to exploit the properties of such OO programs and the natural insights programmers have in their algorithms.

Programs written using the runtime library closely resemble their sequential versions intended to run on a uniprocessor, but for few user-annotations. Users compose programs from computations and data structures amenable to concurrent execution, as they would conventional parallel programs. In addition, they annotate the code to identify *potentially* concurrent functions and the data potentially shared between them. They further formulate the shared data read and written (in the form of objects) by the functions, available from the function's interface, into read and write sets, respectively. Beyond these annotations the onus is not on the user to schedule execution of the computations or to en-

sure independence of concurrent computations, in contrast to conventional parallel programmming.

2.2 Executing Programs

To execute a program on processing cores the runtime raises the granularity of computations to functions. It sequences through the program sequentially but seeks to execute the functions concurrently. Before executing a function the runtime establishes its dependence on already executing functions using the objects in the function's read and write sets. Since objects in the read and write sets may be unknown statically, their identity is established dynamically, at run-time, by dereferencing pointers. The runtime employs dataflow execution since it naturally exposes the innate parallelism between computations. Functions found to be independent are submitted for execution while those that are dependent are "shelved" until their dependences have resolved. The runtime continues to seek work beyond stalled computations, resources permitting, and thus dynamically exploits any available parallelism. Moreover, it ensures that the execution proceeds as per the implied semantics that programmers have come to expect from sequential programs.

The runtime also provisions to handle functions (identified by the user) which do not follow OO principles (e.g., with unknown side effects) by executing them sequentially.

Statically-sequential applications (blackscholes, barneshut, bzip2, dedup, histogram, and reverse index) from standard benchmark suites, developed using the runtime on three stock multicore systems, an 8-thread Intel Nehalem-based machine, a 16-core and a 32-core AMD Opteron-based machines, achieved speedups (harmonic mean) similar to their Pthread versions on the Nehalem machine and over 20% better on the AMD Opteron machines [1].

2.3 Time- and Energy-Efficient Execution

Utilizing resources efficiently in dynamically changing environments will be a key challenge going forward. Doing so will require exposing application parallelism that best fits the capabilities of resources in the execution environment. While exposing too little parallelism can underutilize the resources, exposing excessive parallelism can lead to contention for resources, potentially degrading its time- and energy-efficiency. Dynamically matching the exposed parallelism with the changing capabilities of the execution environment requires the ability to suspend already executing computations, reintroduce them later, and introduce new computations into the environment, as appropriate. The runtime exploits the implied ordering in statically-sequential programs to choose computations judiciously when regulating the parallelism, while ensuring forward progress. It uses a *Goodness of Parallelism (GoP)* metric, computed periodically as the execution unfolds, to correlate the instantaneous efficiency of the program to the instantaneous degree of parallelism. A drop in efficiency causes it to throttle the parallelism to ease contention, while an improvement in efficiency causes it to increase the parallelism to exploit available resources.

Experimental results on a stock 4-core (8-thread) Intel Core i7 2600 (Sandy Bridge) workstation show that our approach achieves up to 50% higher time- and energy-efficiency over the state-of-the-art parallel execution systems across a variety of dynamic operating conditions.

2.4 Precise-Restartable Execution

Future computer systems will present unreliable resources to applications due to exception events, e.g., hardware faults, timing errors caused by aggressive energy management, or due to resource management. To be efficient it will still be desirable to continue executing the interrupted program, possibly at a different time and/or on another system, without discarding all of the completed work. Hence to resume execution in such scenarios the runtime supports *precise-restartability* of parallel programs, analogous to precise-interruptible execution of sequential programs.

The runtime exploits the implied ordering to precisely identify the excepted computation in the statically-sequential program and restores the program state to reflect the sequential execution of the program up to the computation. To do so it tracks the invocation and completion of computation in the implied program order. Further, it checkpoints the state a computation may modify, i.e., its *mod set* (a user-provided set similar to the computation's write set and processed similarly) before its execution. Once the excepting condition is mitigated the program may resume from the excepting computation. The runtime also incrementally checkpoints the program state after each computation successfully completes, using its mod set. This state can be used to spatially or temporally migrate a halted program.

Experiments on a stock 12-core (24-thread) Intel Xeon E5-2420 (Sandy Bridge) workstation show that the runtime can tolerate signficantly higher (proportional to thread-count) exceptions than the conventional approaches. Depending on the application, the support to tolerate aggressive exception rates (e.g., up to 2 every second) incurs performance overheads ranging from 0% to 135% (at 0 faults).

3. CONCLUSION

Parallel programming for multicore-based systems and their dynamically changing operating environments pose significant challenges to everyday programmers in the effort to improve productivity and to achieve error-free, efficient execution of their programs. We presented a model that meets these challenges better than other approaches by using statically-sequential programs and performing their dynamically controlled dataflow execution.

References

[1] G. Gupta and G. S. Sohi. Dataflow execution of sequential imperative programs on multicore architectures. In *MICRO-44*, December 2011.

Efficient, Precise-Restartable Program Execution on Future Multicores

Gagan Gupta, Srinath Sridharan, and Gurindar S. Sohi

Challenges in Future Computing Systems

- **Programmer Productivity**
 - Simplify programming of dynamic and static heterogeneous multicores
 - Enable portable, platform-agnostic programming
- **Efficiency**
 - Optimize energy- and performance-efficiency
 - Adapt to dynamically and continuously changing operating conditions
- **Reliability**
 - Tolerate increasingly unreliable resources
 - Provide differentiated levels of service

Proposed Model: Sequential Programs, Dynamic Parallelization

- **Statically-sequential** programs (using parallel algorithms)
- **Dynamic dataflow** parallel execution
 - Preserving sequential semantics
- **Dynamically controlled** parallelism
- Implemented with a software runtime library (C++)
 - Seamlessly addresses Productivity, Parallel Execution, Efficiency, and Reliability

Programming

- Leverage modern **Object-Oriented** principles
 - Modularity, encapsulation
- Exploit **users' insights** in their algorithms
 - Users identify *potentially* parallel functions, data *potentially* shared between them, and their read and write sets
- Users do not ensure independence between computations, nor orchestrate parallel execution

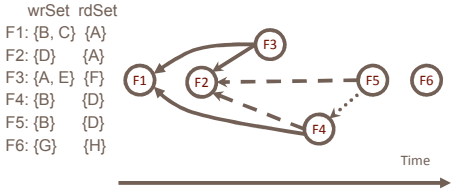
```
1  for (i = 1; i < 7; i++) {
2    df_execute (&F, wrSet, rdSet);
3  }
```

Example of Statically-Sequential Code

- **Simplified parallel programming**

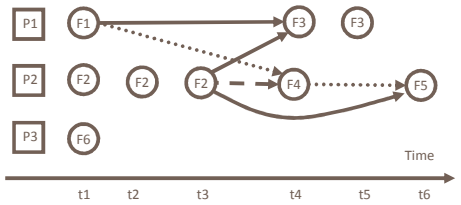
Parallelizing the Execution

- Exploit **Function-Level parallelism**
- Program execution unfolds sequentially
 - Functions execute concurrently (dataflow schedule)
- Data dependences between functions established dynamically (using data sets)

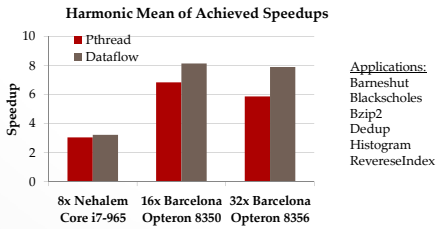


Dynamic Invocations of Example Code and Dependence Graph

- Independent functions execute concurrently, dependent functions are serialized (in program order)
- Dependences are tracked as functions execute and complete



Example Dataflow Execution Schedule on 3 Cores



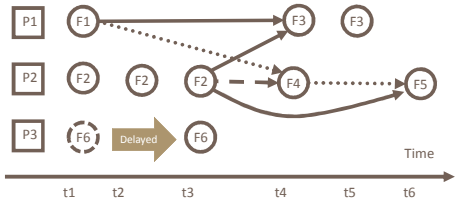
- **Up to 20% higher speedups than conventional implementations**

Benefits of Order and Dataflow Execution

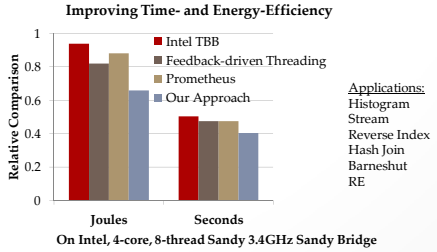
- Dynamically establish dependences to honor
- **Sequentially determinate**, predictable and repeatable execution
- Freedom from deadlocks; guaranteed forward progress
- **Arbitrary control of execution** to optimize efficiency
- **Precise-restartability** of halted computations

Efficient Execution

- "**Goodness of Parallelism**" metric to assess instantaneous efficiency
 - Measured periodically
- **Adapt degree-of-parallelism** to resource contention
 - Optimize for time- and energy-efficiency



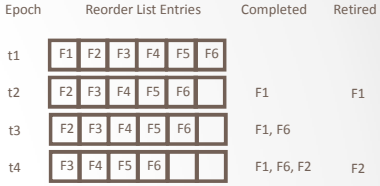
Dynamically Adaptive Execution: F6 Delayed to Improve Efficiency



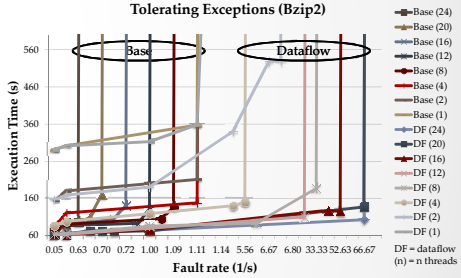
- **Up to 50% more time- and energy-efficient than state-of-the-art parallel execution systems**

Precise-Restartable Execution

- Order of executing computations tracked using a **Reorder List**
- Functions "retired" in program order
- Computation state checkpointed in **History Buffer**
 - Restored on exception, if needed



Precise-Restartable Execution



- **Tolerates significantly higher fault rates (proportional to thread-count) than conventional methods**
- **Incurs 0% to 135% performance overhead (at 0 faults)**

Program Execution on Future Multicores

Dynamically controlled, dataflow execution of statically-sequential programs

- **Enables simplified, platform-independent programming**
- **Automates platform-specific, dynamic and continuous optimizations for energy- and performance -efficiency**
- **Tolerates high fault rates and manages resources at low overheads**